ISCTE IUL
Instituto Universitário de Lisboa

IUL School of Technology and Architecture
Department of Information Science and Technology

Hotel Revenue Management:
Using Data Science to Predict Booking Cancellations

Nuno Miguel da Conceição António

Thesis specially presented for the fulfilment of the degree of

Doctor in Information Science and Technology

Supervisors:
PhD Ana de Almeida, Assistant Professor,
ISCTE-IUL

PhD Luis Nunes, Assistant Professor,
ISCTE-IUL

June, 2019

Instituto Universitário de Lisboa

IUL School of Technology and Architecture
Department of Information Science and Technology

Hotel Revenue Management:
Using Data Science to Predict Booking Cancellations

Nuno Miguel da Conceição António

Thesis specially presented for the fulfilment of the degree of

Doctor in Information Science and Technology

Jury:
PhD Pedro Ramos, Associate Professor, ISCTE-IUL
PhD Markku Vieru, Associate Professor, University of Lapland
PhD Paulo Rita, Full Professor, NOVA IMS (Universidade Nova de Lisboa)
PhD Francisco Serra, Coordinator Professor, Universidade do Algarve
PhD Catarina Silva, Associate Professor, Instituto Politécnico de Leiria
PhD Ana de Almeida, Associate Professor, ISCTE-IUL

June, 2019

To Sofia for all the love and support.

In memory of my mother, Ana Rosa.

ABSTRACT

In the hotel industry, demand forecast accuracy is highly impacted by booking cancellations. Trying to overcome loss, hotels tend to implement restrictive cancellation policies and employ overbooking tactics which in turn reduces the number of bookings and reduces revenue. To tackle the uncertainty arising from cancellations, models for the prediction of a booking's cancellation were developed. Data from hotels' reservations systems was combined with data from other sources (events, holidays, online prices/inventory, social reputation and weather). Despite data class imbalance, concept drift, and dataset shift problems, it was possible to demonstrate that to predict cancellations of bookings is not only possible but also accurate. Moreover, it helped to better understand what the cancellation drivers can be. In order to assess the models under real conditions, a prototype was developed for field tests allowing to evaluate how an automated machine learning system that predicts booking's cancellations could be integrated into hotels' systems. The model's performance in a real environment was assessed, including the impact on the business. The prototype implementation enable an understanding of adjustments to be made in the models so that they could effectively work in a real environment, as well as fostered the creation of a new measure of performance evaluation. The prototype enabled hoteliers to act upon identified bookings and effectively decrease cancellations. Moreover, results confirmed that booking cancellation prediction models can improve demand forecast, allowing hoteliers to understand their net demand, i.e., current demand minus predicted cancellations.

Keywords: Data science, Hotel industry, Machine learning, Predictive analytics, Revenue management

RESUMO

Na indústria hoteleira, a precisão da previsão da procura é altamente impactada pelos cancelamentos de reservas. Na tentativa de mitigar as consequências dos cancelamentos, os hotéis tendem a implementar políticas de cancelamento restritivas e táticas de *overbooking*, o que, por sua vez, reduz o número de reservas e a receita. Para combater a incerteza decorrente dos cancelamentos, foram desenvolvidos modelos capazes de prever a probabilidade de cada reserva vir a ser cancelada. Neste desenvolvimento foram utilizados dados de oito sistemas de gestão de reservas de outros tantos hotéis, conjuntamente com dados de outras fontes (eventos, feriados, preços/inventário *online*, reputação social e clima). Apesar dos problemas de desequilíbrio de classe de dados, desvio de conceito e variação de distribuição entre variáveis ao longo do tempo, foi possível demonstrar que prever cancelamentos de reservas não é apenas possível realizar, mas que é possível de fazer com elevada precisão. A elaboração dos modelos ajudou ainda a compreender os fatores que influenciam o cancelamento. Para avaliar os modelos em condições reais, foi desenvolvido um protótipo, o qual permitiu avaliar como um sistema automatizado baseado em aprendizagem automática para prever os cancelamentos de reservas pode ser integrado nos sistemas dos hotéis. Este protótipo permitiu ainda avaliar o desempenho dos modelos num ambiente real, incluindo o seu impacto na operação. A implementação possibilitou também compreender os ajustes a serem feitos aos modelos para que pudessem efetivamente trabalhar num ambiente real, bem como fomentou a criação de uma nova medida de avaliação de desempenho. O protótipo permitiu que os hoteleiros agissem sobre as reservas identificadas e efetivamente diminuíssem os cancelamentos. Para além disso, os resultados confirmaram que os modelos de previsão de cancelamento de reservas podem melhorar a previsão de procura, permitindo que os hoteleiros compreendam melhor a sua procura líquida, ou seja, a procura atual menos os cancelamentos previstos.

Palavras-chave: Aprendizagem automática, Ciência de dados, Gestão de receita, Indústria hoteleira, Modelos preditivos

ACKNOWLEDGMENTS

Gratitude is perhaps the word that best defines what I feel in completing this long and challenging journey that is to do a doctorate. Although it is a journey that is made up of many periods of solitary work, it is a journey that is impossible to accomplish without the support of other people and institutions. It is good to know that when we need it, there are so many people willing to help. To all of them, and there were many, I want to leave my deepest gratitude here.

Firstly, I would like to express my enormous gratitude to my supervisors, Professor Ana de Almeida and Professor Luis Nunes. Without knowing me, they not only believed in my initial proposal, but they also made an effective contribution in transforming it in a PhD research project and accepted to be my supervisors. Without Professor Ana de Almeida, and Professor Luis Nunes knowledge, guidance, motivation, availability, and patience (I know that sometimes I can be annoying) this project would not have been completed. Their willingness to help in all matters, including in bureaucracies, conferences participations, among other non-research related things, is an example that I will try to follow and something I will never forget.

Besides my supervisors, I would like to thank also the rest of my thesis committee: Professor Paulo Rita and Professor Francisco Serra. Before starting this doctoral course, I did not know Professor Paulo Rita, but during the course, I had the privilege of getting to know him better and beneficiating from his insightful comments and suggestions. Professor Francisco Serra is someone I admire, as a professional and as a person, since he was my MsC supervisor. Although he started working in a public office at the same time I started the doctoral course, he was always available for helping, allowing me to beneficiate from his revenue management and hotel industry expertise.

I would like to thank also all the professors of the classes I had in the doctoral course, at ISCTE-IUL. However, I have to highlight Professors Fernando Batista and Ricardo Ribeiro. It was their classes and subsequent collaboration in two papers, that made me treasure natural language processing.

A special thanks to Professor Paul Phillips from the University of Kent, U.K. The opportunity to work with him was an honor that I cherish and that made me grow as a researcher.

A word of gratitude also to my colleagues at the Universidade do Algarve, in particular, to Marisol Correia, Célia Ramos, Carlos Afonso, José Santos, and Ana Marreiros who have taught me much.

Even though for privacy reasons I cannot mention any names, I would like to express my sincere gratitude to the hotels' board members and hotels' managers who allow me to use their hotels' data in my research and for sharing their valuable time in meetings with me.

Conducting doctoral research and making it known is something that can require some investment. In my case, this was particularly true due to the computational power and data storage

space required. I managed to secure all the infrastructure resources with the help of three different entities. First, ITBASE, the company I work, lent me a server for the duration of the course. Second, a friend, Pedro Reis from Liderlink, lent me two servers on his company datacenter. Third, Microsoft. By conceding me a Microsoft Azure Data Science Grant Award in the value of 50 000 USD it allowed me to overcome all the remaining computational restrictions I had. I also had the financial support of ISCTE-IUL, FCT (funded by FCT/MEC through national funds under the project UID/EEA/50008/2013), ISTAR-IUL, CISUC, Instituto das Telecomunicações, and Fundação Luso-Americana para o Desenvolvimento for sharing and publishing my PhD research. To all of them, my deepest thanks.

A very special gratitude goes to Carlos Soares and Carlos Alexandre Soares that understood what the PhD meant for me and not only agreed on being ITBASE to support the tuition costs but also gave me the latitude to manage my working time according to the PhD requirements.

A big thank you to my friend and fellow diver from PORTISUB, Peter Bolton, for the time he spent revising the English of some of my publications. A big thank you also to my colleague Tatiana Simões for revising some of my many PowerPoint presentations.

I want to thank my family, friends, and fellow members of the board of PORTISUB, for their support and for understanding my absence from many events during this period of the doctoral course.

Last but not least, I would like to thank my wife Sofia. It was several years with virtually no vacations. Many weekends spent working. Many days waking up early or staying up until late to work on the PhD research. Without Sofia's understanding and support it would have been impossible to get to the end of this journey.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

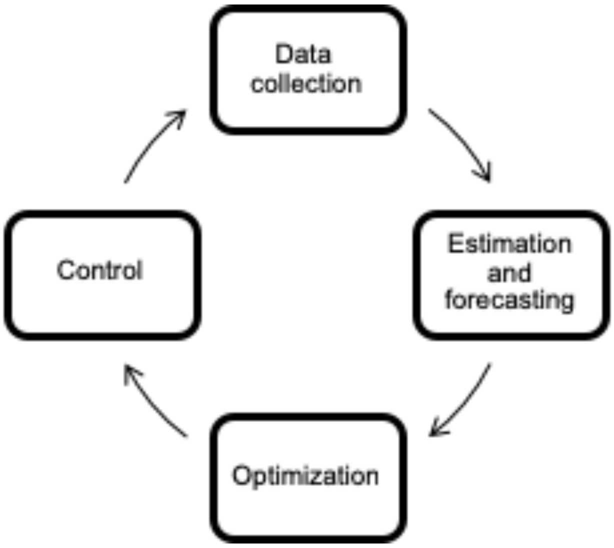
ADR	Average Daily Rate
API	Application Programming Interface
ARC	Airline Reporting Corporation
AUC	Area Under the Curve
BDT	Boosted Decision Tree
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CRS	Central Reservation System
CSV	Comma Separated Values
DJ	Decision Jungle
DM	Data Mining
DSR	Design Science Research
DTM	Document-Term Matrix
ETL	Extract, Transform, and Load
LDSVM	Locally Deep Support Vector Machine
MF	Minimum Frequency
NA	Not Available
NLP	Natural Language Processing
NN	Neural Network
OTA	Online Travel Agency
PMS	Property Management System
PNR	Passenger Name Record
RMS	Revenue Management System
TDM	Term-Document Matrix
TM	Text Mining
TF-IDF	Term Frequency-Inverse Document Frequency
SD	Standard Deviation
SQL	Structured Query Language
WoS	Web of Science

1 INTRODUCTION

Applied mathematics, operational research, machine learning, statistics, databases, data mining, data visualization, and excellent presentation fluency, complemented with a deep understanding of the problem domain are some of the analytic and communication skills in the foundation of data science (Dhar 2013; O'Neil, Schutt 2013; Yangyong, Yun 2011). Despite being quite apart from data science, revenue management tends to be increasingly strategic and technologically oriented. Analytical and communication skills are precisely the type of skills revenue managers have to be better at (Kimes 2010).

Revenue management objectives of increasing revenue and maximizing profitability are achieved through demand-management decisions, i.e., decisions that require the understanding of the characteristics of demand, and its estimation, in order to define prices and capacity controls to manage it (Mehrotra, Ruttley 2006; Talluri, Van Ryzin 2005). Estimation and forecasting is one of the four steps that form the well-known revenue management cyclical process (Figure 1). Therefore, it is not surprising that forecasting performance is a key aspect of revenue management and a critical tool of Revenue Management Systems (RMS). RMS are systems that make use of data, technology and scientific models to automate and assist human analysts in making demand-management decisions (Ivanov, Zhechev 2012; Talluri, Van Ryzin 2005). Without accurate forecasts, RMS rates and inventory availability recommendations would probably be highly inaccurate (Chiang, Chen, Xu 2007; Ivanov, Zhechev 2012; Lemke, Riedel, Gabrys 2013; Mehrotra, Ruttley 2006; Talluri, Van Ryzin 2005; Weatherford, Kimes 2003).

Figure 1 - Revenue management process



Adapted from Talluri, Van Ryzin (2005)

Booking cancellations, together with room nights' occupancy, arrivals, and price sensitivity, are some of the subjects of forecasts in hotel revenue management. Forecasting bookings cancellation is of significant importance to determine net demand, i.e., demand deducted from predicted cancellations and no-shows (Lemke, Riedel, Gabrys 2013; Talluri, Van Ryzin 2005). Cancellations are reservations that are terminated by customers prior to the time of service. No-shows, in contrast, are reservations that fail to show up at the time of service without giving any prior notice (Talluri, Van Ryzin 2005).

Even though forecasting and prediction are considered synonyms and employed interchangeably (Clements, Hendry 1998; Matsuo 2003), scientifically, these words assume different meanings and definitions. While forecasting aims to calculate or predict future events, commonly associated with a time series, prediction can also be used to reconstruct and explain some past outcome (Lewis-Beck 2005; Matsuo 2003). By solely recurring to statistical modeling for causal explanation and not complementing with predictive modeling, some fields are neglecting the relevance of existing theories and the ability to discover new causal mechanisms. Moreover, these fields are not seizing the opportunity to combine explanatory and predictive modeling to bridge the gap between methodological development and practical application (Shmueli 2010). This difference in terminology is the reason the term prediction is used in the title of this work instead of forecasting, a term commonly employed in revenue management literature. However, it should be clarified that in revenue management research some authors, like Talluri, Van Ryzin (2005), employ the term estimation as a synonym for prediction, in particular when describing models built not only to forecast what will be observed, but also what has been observed.

Although this dissertation lays in the discipline of information science and technology, being a specialization on the application of information technology to management and social sciences, makes it a cross-disciplinary work. As such, to ease the comprehension for audiences from both information science and hotel revenue management disciplines, some basic concepts in the respective disciplines will be introduced whenever necessary.

The first two sections of this initial chapter present a description of the background, motivation, aim, and scope of this dissertation. The following section enumerates the expected contribution of this dissertation to theory, methodology, and practice. A list of publications that directly and indirectly resulted from the research conducted for this dissertation will be next presented and one other section with an overview of the dissertation structure concludes the chapter.

1.1 Background and motivation

Revenue management is defined as *"the application of information systems and pricing strategies to allocate the right capacity to the right customer at the right price at the right time"* (Kimes, Wirtz 2003, p. 125). The term "revenue management" is usually employed to describe the broad variety of techniques, methods, processes, and technologies involved in making demand-management decisions. Typically, demand-management decisions are grouped in three categories (Talluri, Van Ryzin 2005):

- Structural: defining the selling format (e.g., posted prices or auctions), segmentation of differentiation mechanisms, terms of trade (e.g., volume discounts, cancellation or refund options);
- Price: defining standard prices and special prices, prices across product/service categories, price overtime, discounts to practice over the product/service lifetime;
- Quantity: defining which offers to buy accept or reject, the capacity of each product/service to allocate to the different segments and distribution channels, when to withhold a product/service from the market and when to sale it in a later point in time.

Demand-management decisions are most effective when applied in industries with fixed capacity, variable and uncertain demand, perishable inventory, high fixed costs and variable customer price sensitivity (Kimes, Wirtz 2003; Talluri, Van Ryzin 2005). Since its development in 1966 by the airline industry (Chiang, Chen, Xu 2007), the application of revenue management extended to other industries such as hotels, rental cars, golf courses, cruise ships, and casinos (Chiang, Chen, Xu 2007; Kimes, Wirtz 2003; Phillips 2005; Talluri, Van Ryzin 2005). The main contribution of revenue management to these industries was not the innovation on demand-management decisions by themselves, but the innovation in the method for decision making – a method highly sophisticated, detailed, intensely operational, based on science and technology. Scientific advances in statistics, operations research and economics made it possible to model demand, produce improved predictions and forecasts and compute optimal solutions to complex decision problems. Information technology advances made it possible to automate transactions, gather large amounts of data and execute complex algorithms. This combination of advances in science and technology made it possible to manage demand in a scale and complexity not possible until then (McGuire 2017; Talluri, Van Ryzin 2005; Phillips 2005).

Although airline, railway, hotel, and other travel related service industries have many similarities and share many of the characteristics required for the application of revenue management, they also have many dissimilarities. Though some chains own many hotels, sometimes even using multiple brands, there are still many independent hotel owners. Hotels can have many differences between them, whether in type (business, resorts or a mix), size (large or small), or location (airport, urban, beach, among others). Some hotels have significant secondary sources of revenue, like food and beverage, function space, or activities (e.g., golf, ski, or gambling). Hotels can have many room types and room rates but, unlike airlines, hotels do not tend to have prices for multi-resource inventory (staying of multiple nights) (Talluri, Van Ryzin 2005) or to use advanced purchase discounts as a segmentation mechanism (Talluri, Van Ryzin 2005). While in the airline industry factors like price, onboard service quality, or safety record can determine the selection of service provider, in hotels other factors like location, social reputation, or duration of stay can be more important (Al Saleem, Al-Juboori 2013; Anderson 2012; Chen, Tsai, Chiu 2017; Jones, Chen 2011). Likewise, capacity in airlines and railways is not as fixed as in hotels. As an example, for high-demand special situations, the airline company can change the model of the airplane to a model with larger capacity, or the railway company can add additional wagons to increase capacity. This hotel industry fragmentation and differentiation of type, size, location, sources of revenue, pricing structure, capacity stiffness, and of other factors that affect service provider selection show the difference between the application of revenue management practices in the hotel industry versus other travel related service industries (Talluri, Van Ryzin 2005).

Hotels use bookings as an essential component to match demand with capacity. Because hotels have limited capacity, customers must book in advance in order to ensure availability (Chen, Xie 2013; Lee 2018; Smith, Parsa, Bujisic, van der Rest 2015). Consequently, bookings (also known as reservations or advanced reservations), are considered to be the primary predictor of a hotel's forecast performance (Lee 2018; Smith, Parsa, Bujisic, van der Rest 2015). A booking represents a forward contract between the customer and the hotel. This contract gives the customer the right to use the service in the future at a settled price, but often with an option to cancel prior to service provision (Talluri, Van Ryzin 2005).

Booking cancellations occur for multiple reasons. Sometimes, due to understandable reasons like illness, bad weather, business meetings, calendar changes or vacation rescheduling. Other times, the reasons behind cancellations are unclear. There is increasing evidence that a significant part of cancellations is made by "deal-seeking" customers, that is, customers who keep searching for the same or similar product/service at a lower cost after booking (Chen, Xie 2013; Chen, Schwartz, Vargas 2011). In some situations, to preserve their options, costumers enter multiple bookings and then cancel all except one (Talluri, Van Ryzin 2005). Hence, it is understandable that customers value the option to cancel bookings. In fact, *"a cancellation option gives customers the best of both worlds - the benefit of locking-in availability in advance and the flexibility to renege should their plans or preferences change"* (Talluri, Van Ryzin 2005, p. 130).

However, the option to cancel places a two-sided risk on hotels. Hotels have to honor bookings and have available rooms for customers that show up but, at the same time, have to bear the opportunity cost of ending up with vacant rooms when a customer cancels or does not show up (Smith, Parsa, Bujisic, van der Rest 2015; Talluri, Van Ryzin 2005). To mitigate the risks associated to cancellations and no-shows, hotels implement a combination of overbooking and cancellation policies (Chen 2016; Chen, Schwartz, Vargas 2011; Ivanov 2014; Mehrotra, Ruttley 2006; Noone, Lee 2011; Phillips 2005; Smith, Parsa, Bujisic, van der Rest 2015; Talluri, Van Ryzin 2005). Overbooking, i.e., accepting more bookings than the physical capacity, has an essential economic role in hotels as a way to address the loss of revenue associated to cancellations, no-shows and early departures (Noone, Lee 2011; Phillips 2005; Talluri, Van Ryzin 2005). Similarly, cancellation policies, in particular, restrictive cancellation policies, by requiring a fee to secure bookings or compensation in case of cancellation or a no-show, not only limit the loss of revenue but also are a non-negligible source of revenue. Moreover, restrictive cancellation policies can also have the positive effect of conditioning the search and booking behavior of “deal-seeking” customers (Benítez-Aurioles 2018; Chen, Schwartz, Vargas 2011; Chen, Xie 2013; Xie, Gerstner 2007).

Nevertheless, overbooking and restrictive cancellation policies can also be detrimental to hotels. Customers who book in advance expect to use the service. At the same time, customers who book and then cancel do not expect to pay since, from their perspective, they did not use the service (Talluri, Van Ryzin 2005). The denial of service provision to customers due to capacity overselling can have a terrible effect on customers, leading to complaints and generating a negative impact in social reputation and brand image (Guo, Dong, Ling 2016; Ivanov 2014; Wirtz, Kimes, Theng, Patterson 2003). Another adverse effect is the loss of immediate revenue directly associated to reallocation and compensation costs (Ivanov 2014; Noone, Lee 2011). Moreover, the possible loss of future revenue is almost inevitable since dissatisfied customers may not book again at the same hotel or at the same brand (Mehrotra, Ruttley 2006; Wirtz, Kimes, Theng, Patterson 2003). In turn, restrictive cancellation policies, in particular, non-refundable or those with cancellation deadline superior to 48 hours, may lead to a decrease, both in the number of bookings and in revenue. The former due to customer’s lower propensity to book when rigid policies apply. The latter due to the discounts associated to non-refundable and rigid policies (Chen, Schwartz, Vargas 2011; Park, Jang 2014; Smith, Parsa, Bujisic, van der Rest 2015).

Balancing the positive and negative effects of overbooking and restrictive cancellation policies is difficult (DeKay, Yates, Toh 2004; Ivanov 2014; Phillips 2005; Smith, Parsa, Bujisic, van der Rest 2015). This difficulty may explain hotels’ high cancellation rates, which can vary between 20% to 60% (Liu 2004; Morales, Wang 2010) and why hotels are less reluctant to ask for a full payment at the time of booking or impose higher cancellation fees when compared to other travel related industries (Smith, Parsa, Bujisic, van der Rest 2015; Chen 2016). Thereby, several authors advocate the development of booking cancellation forecast and prediction models in order to improve demand forecast in revenue management (Chen 2016; Hueglin, Vannotti 2001; Lemke, Riedel, Gabrys 2013; Morales, Wang 2010; Talluri, Van Ryzin 2005). Research in this topic is still scarce, in particular for the hotel industry (Benítez-Aurioles 2018; Chen 2016). Most of the existing research relates to airlines and relies on a single data source (Iliescu, Garrow, Parker 2008; Lemke, Riedel, Gabrys 2013; Petraru 2016). Furthermore, the existing literature employ time series historical aggregated data or detailed booking data in the Passenger Name Record (PNR) format, a standard created by the airline industry and for the airline industry (International Civil Aviation Organization 2010). However, the use of industry-specific data sources, like data from hotels' Property Management Systems (PMS), or the use of data from other sources, like weather forecast data, events data, or macroeconomic data, could possibly improve forecast accuracy (Chiang, Chen, Xu 2007; McGuire 2017; Ivanov, Zhechev 2012; Pan, Yang 2017a; Talluri, Van Ryzin 2005).

1.2 Aim and scope

This dissertation aims to improve hotels' net demand forecasting and decrease uncertainty in demand-management decisions by demonstrating that using data science it is possible to build models to predict bookings' cancellation likelihood and to understand cancellation drivers.

Although there are similarities between travel related service industries and hotels, the characteristics of each industry prevent many of the models already employed in revenue management to be generalized across industries. Not, at least, without changes and adaptations (Phillips 2005; Talluri, Van Ryzin 2005). Moreover, because the author professional activity is related to the hotel industry, there is a particular interest in the scientific developments for this industry. This relation with the hotel industry allows the author to have access to real-world hotel data, which is something that is always required for developing and testing predictive models. For these reasons, and mainly because the subject of bookings cancellation forecast/prediction is a subject understudied in the hotel industry, it was decided to limit the scope of this dissertation to this industry.

The development of models for the hotel industry with the sole purpose of forecasting/predicting no-shows is not as significant as in other industries (Lawrence, Hong, Cherrier 2003; Neuling, Riedel, Kalka 2004; Zenkert 2017). Therefore, estimating the probability of a customer not showing up at the time of service is not considered relevant in the scope of the problem addressed in this dissertation. As such, building a static model to predict each booking's likelihood of canceling, does not require a distinction between no-shows and cancellations (Talluri, Van Ryzin 2005).

1.3 Contributions

The primary outcome of this dissertation is to show that combining hotel specific data (PMS data) with data from other sources, and applying data science tools and capabilities, such as advanced machine learning classification algorithms, it is possible to build hotel bookings' cancellation prediction models that can predict, with high accuracy, the likelihood of each booking to be canceled.

At a theoretical level, the models may allow understanding which features[1] drive cancellations. This understanding, in turn, can contribute to increase knowledge on how to define more balanced cancellation policies, that is, policies not as rigid to the point of pushing away customers but, at the same time, give hotels some guarantees for revenue if cancellations do occur.

From a methodological point of view, the contributions of this dissertation are centered around data sources' types and model elaboration. The fact that showing what data sources should be employed, what features should be used, how to engineer features, how to deal with the typical problems for this type of datasets, how to employ the classification algorithms for these problems, and how these models could be implemented in an RMS, makes this dissertation contributes to the advance of knowledge on how to build and deploy booking cancellation prediction models. This knowledge could be applied by researchers in other hotel industry prediction problems or adapted to other travel industries prediction problems.

On a practical level, the predicted outcome of the models could be used to calculate net demand at an aggregated level (global demand) or a more detailed level (e.g., per market segment or distribution channel). Net demand could be used as a metric to make better overbooking decisions and with that mitigate the reallocation, compensation and reputation costs associated to "walked"[2] customers.

The prediction outcomes of the models can be used to identify which customers should a hotel contact prior to arrival in order to take actions to prevent a potential cancellation. Understand how customers react to this contact, in other words, understand if customers tend to cancel more or cancel less when contacted by hotels is also an important contribution of this dissertation.

During the development of this dissertation, partial results and findings were published in peer-reviewed books, journals and conferences. These publications are summarized in the following section.

1.4 List of publications

The list of publications that resulted from the research for the elaboration of the dissertation is divided in two subsections: one for publications directly related to the dissertation theme and another with publications indirectly related to it. In total, sixteen publications were made. From the sixteen, twelve are already published and four are in revision or accepted for presentation in conferences. From the twelve published, six are papers published in Scopus and Web of Science

[1] The term "feature" in machine learning has the same meaning as the term "variable" or "independent variable" in traditional statistics, so it is common for terms to be used interchangeably. However, "feature" is used in this work rather than "variable" because it is frequent to replace variables by a computational result from one or more input variables.

[2] "Walked" customer is a term used in the hotel industry to designate customers that, due to overbooking, have to be walked (reallocated) to another hotel.

(Wos) journals (three of which Scopus Q1 in 2018), five are papers presented in renown conferences proceedings (one of which is index in Scopus), and one is a book chapter (also indexed on Scopus).

1.4.1 Directly related

List of publications directly related to the dissertation work:

1. ANTONIO, Nuno, ALMEIDA, Ana de and NUNES, Luis, 2016. Predicting hotel booking cancellation to decrease uncertainty and increase revenue. In: Book of Abstracts of TMS Algarve 2016. Olhão, Portugal: ESGHT, Universidade do Algarve. November 2016. p. 47. ISBN 978-989-8472-93-9.
2. ANTONIO, Nuno, ALMEIDA, Ana de and NUNES, Luis, 2017. Using data science to predict hotel booking cancellations. In: Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry. Hershey, PA, USA: Business Science Reference. p. 141–167. ISBN 978-1-5225-1054-3.
3. ANTONIO, Nuno, ALMEIDA, Ana and NUNES, Luis, 2017. Predicting hotel booking cancellation to decrease uncertainty and increase revenue. Tourism & Management Studies. Vol. 13, no. 2, p. 25–39. DOI 10.18089/tms.2017.13203.
4. ANTONIO, Nuno, ALMEIDA, Ana de and NUNES, Luis, 2017. Enabling bookings cancellation with data science. In: Book of Abstracts of the 4th World Research Summit for Tourism and Hospitality. Orlando, FL, USA: Elsevier/UFC Rosen College of Hospitality Management. 1 December 2017.
5. ANTONIO, Nuno, ALMEIDA, Ana de and NUNES, Luis, 2018. Predicting hotel bookings cancellation with a machine learning classification model. In: Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017. Cancun, Mexico: Institute of Electrical and Electronics Engineers Inc. January 2018. p. 1049–1054. ISBN 978-1-5386-1417-4.
6. ANTONIO, Nuno, ALMEIDA, Ana de and NUNES, Luis, 2018. Predictive models for hotel booking cancellation: a semiautomated analysis of the literature. In: Book of Abstracts of TMS Algarve 2018. Olhão, Portugal: ESGHT, Universidade do Algarve. November 2018. p. 33. ISBN 978-989-8859-53-2.
7. ANTONIO, Nuno, ALMEIDA, Ana and NUNES, Luis, 2019. Predictive models for hotel booking cancellation: a semi-automated analysis of the literature. Tourism & Management Studies. Tourism & Management Studies. Vol. 15, no. 1.
8. ANTONIO, Nuno, ALMEIDA, Ana and NUNES, Luis, 2019. Hotel booking demand datasets. Data in Brief. Vol. 22, February 2019, p. 41-49. DOI 10.1016/j.dib.2018.11.126.
9. ANTONIO, Nuno, ALMEIDA, Ana de, NUNES, Luis. Big data in the hospitality industry: Exploring cancellation drivers to gain insights into booking cancellation behavior. Cornell Hospitality Quarterly. DOI 10.1177/1938965519851466.

10. ANTONIO, Nuno, ALMEIDA, Ana de, NUNES, Luis, 2019. An automated machine learning decision support system to predict hotel booking cancellations. Data Science Journal (accepted for publication).

Additional information, including publication indexation, brief description, and the chapters covered is detailed in Table 1.

Table 1 - Directly related documents additional information

ID	Main indexations	Publisher	Brief description	Chapter
1	-	ESGHT, Univ. do Algarve	Uses four hotel' bookings datasets for data exploration and building of exploratory predictive models	3
2	Scopus	IGI Global	Uses four hotel' bookings datasets for data exploration and building of exploratory predictive models	3
3	Web of Science, Scielo	ESGHT, Univ. do Algarve	Extends the work described in publication with ID 2, using different methods for the development of the models	3
4	-	Elsevier/UFC Rosen College	Describes the prototype development and preliminary results	4 and 5
5	Scopus	IEEE	Describes the prototype development and preliminary results	4 and 5
6	-	ESGHT, Univ. do Algarve	Uses text mining and natural language processing for a semiautomated literature review about booking's cancellation prediction	2
7	Web of Science, Scielo	ESGHT, Univ. do Algarve	Uses text mining and natural language processing for a semiautomated literature review about booking's cancellation prediction	2
8	Scopus	Elsevier	Open share of the datasets employed in the prototype's development and deployment	5
9	Scopus, Web of Science	SAGE	Describes the development of the final models, including the process to extract and integrate data from the different data sources	4
10	Scopus	Ubiquity	Describes in detail the process for building and implementing the prototype, including the final results of its use	5

1.4.2 Indirectly related

The ISCTE-IUL PhD program included mandatory and optional courses. One of the optional courses was Natural Language Processing (NLP). This course led to a participation in the 2016 edition of the Lisbon Machine Learning School, a summer school targeted to researchers and graduate students in the fields of NLP, computational linguistics and machine learning. During the courses the author used NLP, machine learning and other data science tools to analyze and explore social reputation data which was extracted for the research on the dissertation theme. The results were interesting enough to proceed with their publication. Subsequently, other colleagues became interested in the work, which was been spawned by a number of collaborations. A list of the publications that resulted from the NLP course and subsequent collaborations is next presented:

1. ANTONIO, Nuno, ALMEIDA, Ana de, NUNES, Luis, BATISTA, Fernando and RIBEIRO, Ricardo, 2018. Hotel online reviews: different languages, different opinions. Information Technology & Tourism. April 2018. Vol. 18, no. 1–4, p. 157–185. DOI 10.1007/s40558-018-0107-x.
2. ANTONIO, Nuno, ALMEIDA, Ana de, NUNES, Luis, BATISTA, Fernando and RIBEIRO, Ricardo, 2018. Hotel online reviews: Creating a multi-source aggregated index. International Journal of Contemporary Hospitality Management. October 2018. Vol. 30, no. 10, p. 3574-3591. DOI 10.1108/IJCHM-05-2017-0302.
3. ROCHA, Érica, SOUSA, Paulo Faria de, ANTONIO, Nuno, MEDEIROS, Susana and JULIÃO, Miguel, 2018. O conceito de dignidade em idosos não institucionalizados a residir em Portugal - Modelo empírico. In: Livro de resumos do 35o Encontro Nacional da Associação Portuguesa de Medicina Geral e Familiar. Vilamoura, Portugal: Associação Portuguesa de Medicina Geral e Familiar. 14 March 2018.
4. ROCHA, Érica, SOUSA, Paulo Faria de, ANTONIO, Nuno, MEDEIROS, Susana and JULIÃO, Miguel, 2018b. O Conceito de dignidade em Idosos não-institucionalizados seguidos em cuidados de saúde primários: Um modelo empírico preliminar. *Acta Médica Portuguesa*. August 2018. Vol. 31, no. 13, p. 1–2. DOI 10.20344/amp.10943.
5. PHILLIPS, Paul, ANTONIO, Nuno, ALMEIDA, Ana de, NUNES, Luis, 2019. The influence of geographic and psychic distance on online hotel ratings. Journal of Travel Research (accepted for publication).
6. RIBEIRO, Filipa Perdigão, CORREIA, Marisol B., ANTONIO, Nuno. Uma abordagem metodológica para a análise comparativa de comentários de viagens online de duas cidades património da Unesco. In: Anais da Conferência Internacional Turismo & História – TURHIST. Faro, Portugal: Universidade do Algarve (Portugal) and Universidade de Caxias do Sul (Brasil).

In addition to these publications, four other publications (two papers and two book chapters) are also under development, in collaboration with thesis supervisors and other authors.

1.5 Thesis structure

The background information provided in this introductory chapter is extended in the literature review and discussion presented in Chapter 2. Even though this dissertation is about the hospitality industry, the literature review covers literature about other travel industries to understand if the models or processes there developed could be of interest to be used or adapted for hotels. In the first moment, the overall importance of forecasting and prediction in revenue management is explained and analyzed. A detailed analysis of the state of the art in booking cancellation forecasting and prediction modeling follows.

Chapter 3 explores PMS data from four resort hotels to perceive how the data can be used to develop bookings cancellation models, analyzes its performance when compared to models using data in the standard PNR format, and identifies its limitations.

Chapter 4 shows how the limitations found in Chapter 3 can be overcome with models that employ exclusively PMS data and models that employ data from multiple sources. Models developed in this chapter employed data from 8 resort and city hotels, together with data from additional sources, such as weather forecast, competitive social reputation, events in the hotel region, among others.

Chapter 5 explores how booking cancellation prediction models can be integrated into RMS and implemented in a real production environment. This evaluation was achieved through the development of a prototype that was in production and under inspection throughout three months in two of the hotels. The results of the models' performance and A/B tests conducted during the tests are presented and discussed.

Chapter 6 summarizes results, contributions, and implications for hospitality revenue management research and practitioners. Additionally, this chapter also summarizes the limitations and directions for future research.

2 Related work and objectives

Bookings cancellation forecasting and prediction, as introduced in the previous chapter, is of critical importance for RMS' forecasting performance. However, as also introduced in the previous chapter, research on this topic is scarce, in particular for the hospitality industry. Although there is literature related to the railway industry, restaurant, and other travel-related industries, the existing literature is predominantly related to the airline industry. Given this scarcity, section 2.1, presents a review of relevant literature on the topic of bookings cancellation forecast and prediction related to travel service industries, focusing in industries that share some hotels' characteristics, like being reservation-based and have a relatively fixed capacity. With this, advantages and disadvantages of the different approaches are shown in order to identify how this dissertation can fill these disadvantages, identify if it is possible to adapt models from other industries to the hotel industry, and also identify areas where this is space for improvement. Based on this critical review of relevant literature, section 2.2 presents an overall discussion of the gaps found in literature and on future research directions to introduce the main questions this dissertation will attempt to answer. Section 2.3 concludes with a summary of what has been discussed in this chapter.

2.1 Literature review

Webster, Watson (2002) consider that there are two types of literature reviews: one regarding a mature topic that has an extended body of knowledge, requiring analysis and synthesis; and another on a promising topic which could benefit from further theoretical groundwork. This chapter will address both. Section 2.1.1 will address the former type. It will provide an analysis and synthesis of forecasting and prediction research in revenue management. This analysis was made through a systematic review of the literature using content analysis as the main method. Section 2.1.2 will address the latter type of literature review. This section provides an analysis of the state of the art in booking cancellation models, including the techniques employed, performance accuracy, and industries where models are or can be applied. This analysis was made through a semiautomated review literature. Because this is a relatively novel method, methodologic steps will be described in detail. In both sections, only papers published between 1990 and May 2018 were considered for analysis. Papers that resulted from this dissertation were removed from the analysis in order not to bias the analysis. Nevertheless, some of the papers studied cited some of the publications that resulted from this dissertation work, which can also influence their authors' comprehension of the problem (Benítez-Aurioles 2018; van Leeuwen 2018).

2.1.1 Forecast and prediction in revenue management

This section complements the brief introduction to revenue management presented in the previous chapter, starting with a succinct retrospective of its application, followed by a description of its major components, continuing with a presentation of forecasting/prediction examples and methods employed, and finalizing with what research say about what should be the future of forecasting and prediction modeling in revenue management.

Revenue management is considered one of the most successful application areas of operations research. Although revenue management started being applied by some airlines in the 1960s, it was not until 1978, with the USA Airline Deregulation Act, that it become widely-adopted by the airline industry. It was not until the mid-1980's that other service industries started to adopt it (Chiang, Chen, Xu 2007; Denizci Guillet, Mohammed 2015; Talluri, Van Ryzin 2005). Revenue management is usually applied in industries where tactical demand management is essential (Talluri, Van Ryzin 2005). However, demand management is complex. The main reason for this complexity is the multiple dimensions that affect demand: (1) type of products/services sold by a company, (2) type of customers, their preferences and purchase comportments, (3) time, (4) location, (5) distribution channels, among others (Phillips 2005; Talluri, Van Ryzin 2005). This multidimensionality explains, for example, why customers attribute different value to a product/service at different moments in time (Denizci Guillet, Mohammed 2015; Phillips 2005; Talluri, Van Ryzin 2005). Revenue management exploits this multidimensional landscape by making structural, price, timing and quantity decisions in order to improve revenue and maximize profitability (Phillips 2005; Talluri, Van Ryzin 2005). These decisions involve cycling through a four-step process at repeated intervals in time (see Figure 1).

Data collection is the first step in the revenue management process. Without relevant historical data (prices, demand, and other factors) estimation and forecasting would not be possible. Historical data is required to estimate forecasting parameters and to forecast quantities like no-shows or cancellations. These parameters and quantities are then used as inputs in demand forecasting models. Based on demand forecasts, the optimal set of controls is defined (inventory allocations, prices, discounts, overbooking limits, etc.). Control of this optimization is then made through monitoring inventory sales across the different transaction systems (Talluri, Van Ryzin 2005).

Given that forecasting is considered one of the five areas of revenue management problems (the others are pricing, auctions, capacity control, and overbooking) (Chiang, Chen, Xu 2007), it is not surprising that forecasting is a topic addressed by a large proportion of revenue management publications (Ivanov, Zhechev 2012). In a survey on the use of forecasting models in revenue management, Weatherford (2016) found that in the period between 1958 and 2016, 83 articles were published on the subject. However, from these, only six were specific to hotel demand forecasting. In another literature review on the topic of revenue management research in hospitality and tourism, Denizci Guillet, Mohammed (2015) identified that from a total of 158 studies published from 2004 to 2013, ten were about demand forecasting. After pricing, customer and distribution channel management, demand forecasting was one of revenue management research dominant topics.

In the airline industry, the literature shows that a 10% increase on forecast accuracy can translate in up to 3% increase in revenue (Lee 1990). It is therefore comprehensible why forecast accuracy is considered of foremost importance to revenue management (Chiang, Chen, Xu 2007; Lee 2018; Talluri, Van Ryzin 2005; Weatherford 2016), and why forecasting consumes a significant part of an RMS development, maintenance, and implementation time (Talluri, Van Ryzin 2005). Demand forecasting is not the only type of prediction or forecasting made by RMS. Many characteristics of demand must also be predicted and forecasted, for example: booking-curve (how demand evolves over time), booking-profile (who will book and when), no-shows and cancellations, revenue, price sensitivity, length-of-stay, cross-selling and up-selling probabilities, competition behavior, amendments to bookings, among other operational metrics (Ivanov, Zhechev 2012; Talluri, Van Ryzin 2005; Weatherford 2016). These characteristics are often used as inputs in overall demand forecasting and optimization (Phillips 2005; Talluri, Van Ryzin 2005; Weatherford 2016).

Based on Lee (1990), Ivanov, Zhechev (2012) and Weatherford, Kimes (2003) categorized forecasting methods in historical, advanced booking and combined. Historical methods are based on traditional forecasting methods such as the various forms of exponential smoothing (e.g., simple or weighted moving average), time series, or linear regression. Advanced booking methods use the number of reservations on hand to forecast future bookings. These methods are divided into additive (e.g., classical or advanced pickup), multiplicative (e.g., synthetic booking curve), or other time series. Combined methods use a combination of historical and advanced booking methods. Before 2000, traditional forecasting methods, mostly based on time-series methods and historical time series data where the only type of methods and data used in revenue management demand forecasting (Pereira 2016; Weatherford 2016). Technology advances, namely at the level of processing power, big data, and artificial intelligence have facilitated the development of new forecast/prediction methods and algorithms that enable the resolution of larger and complex mathematical problems. Artificial intelligence, in particular, machine learning models, are models built on a set of test data and deployed on unknown data that to perform the same task. Logistic regression, clustering, decision trees, and neural networks are some of the algorithms classified as machine learning algorithms (McGuire 2017). Although there is some evidence of the application of machine learning methods and algorithms in travel-related service industries to solve revenue management problems (McGuire 2017), the topic is understudied in the scientific literature. Most of the few examples found in literature explore the application of neural networks (Freisleben, Gleichmann 1993; Law 2000; Huang, Chang, Ho 2013; Padhi, Aggarwal 2011; Weatherford, Gentry, Wilamowski 2003; Zakhary, Gayar, Ahmed 2010). Other examples explore the use of algorithms such as decision trees, support vector machine, logistic regression, or Naïve Bayes (Hueglin, Vannotti 2001; Lawrence 2003; Morales, Wang 2010; Neuling, Riedel, Kalka 2004).

Besides the differences in quantities or measures forecasted/predicted and the differences in methods employed, forecasts and predictions can also be distinguished by the level of aggregation (Talluri, Van Ryzin 2005; Weatherford 2016). Depending on what is the subject of the forecast and the level of data detail (the more desegregated the forecast required, the more detailed the data must be), one of two strategies is followed: “bottom-up” or “top-down” (Talluri, Van Ryzin 2005; Weatherford, Kimes, Scott 2001). A “bottom-up” strategy is used when detailed forecasts are required (e.g., occupancy per room type, per night). Forecasts can then be added up to obtain global results (e.g., overall occupancy, per night). A “top-down” strategy is used to make global forecasts, which results can then be used to disaggregated forecasts (e.g. use a global forecast of customers per rate category to forecast the length of stay of customers).

One other characteristic that distinguishes the type of forecasts and prediction problems is the type of the target variable. From a machine learning point of view, supervised forecast and prediction problems should be categorized as regression problems when the target variable is continuous or, as classification problems when the target variable is categorical (Abbott 2014; Hastie, Tibshirani, Friedman 2001).

Despite the dynamic nature of revenue management, the size of revenue management problems, the limitations of reservation systems, and the emergence of new business models, forecasting is still considered to be difficult, costly and occasionally failing to produce satisfactory results (Weatherford 2016). Forecasting models that worked well in the past may not work well in the future (Chiang, Chen, Xu 2007). Additionally, the fact that some forecasting models remain proprietary knowledge of some hotel chains also difficult the theoretical advance of the field (Ivanov, Zhechev 2012). Hence, literature urges future research to take advantage of technological and mathematical/scientific methods, including big data and machine learning, to develop new and improved forecasting models (Chiang, Chen, Xu 2007; Ivanov, Zhechev 2012; McGuire 2017; Pan, Yang 2017a; Weatherford 2016).

Next section delves further in the level of detail, presenting the state of the art on models specifically developed to forecast or predict cancellations.

2.1.2 Bookings cancellation forecast and prediction

This section investigates the state of the art of bookings cancellation forecast and prediction models in revenue management, in the scope of demand forecasting for hotel and other related reservation-based travel industries. Data science tools and capabilities, especially Text Mining (TM), NLP and data visualization, are employed to conduct a semiautomated analysis of existent literature. The section starts with an introduction to automated literature review, followed by a detailed presentation of the methodology, and concludes with the results presentation.

2.1.2.1 Automated literature review

Conducting comprehensive literature reviews is becoming increasingly complex. The increasing abundance of potentially relevant research, not only on the research field but also in related and even non-related fields, makes the task evermore demanding (Delen, Crossland 2008; Nunez-Mir, Iannone, Pijanowski, Kong, Fei 2016). This progressive difficulty in carrying out an adequate literature review causes some authors, at least in some fields, to defend the need to take advantage of technological advances to automate literature reviews. This automation would potentially enable faster and less resource-intensive literature reviews (Bragge, Relander, Sunikka, Mannonen 2007; Feng, Chiam, Lo 2017; Tsafnat, Glasziou, Choong, Dunn, Galgani, Coiera 2014). In fact, in a literature review conducted by Feng, Chiam, Lo (2017), the authors found 32 relevant studies that advocated the use of automated or semiautomated solutions to support systematic literature reviews.

As identified by Tsafnat, Glasziou, Choong, Dunn, Galgani, Coiera (2014), the automation of literature review has the potential to help researchers in almost all of systematic review tasks, namely: formulation of the review questions, finding previous systematic reviews, writing the protocol, devising the search strategy, searching, finding duplicates, scanning abstracts, obtaining full text articles, scanning full text articles, forward and backward citation searching, data extraction, data conversion and synthetization, literature re-checking and lastly, writing up the review. Delen, Crossland (2008) state that the application of advanced methods that allow for the automation of literature review could potentially lead to: enhancement of the retrieval of data, characterization of the research based on metadata (journal, authors, organizations), to reveal new technical concepts or technical relationships, identification of the main topics and sub-topics of research, identification of the relationship between topics and metadata, provision of insights on research directions. The application of these advanced methods and the benefits can be witnessed by some examples of automated, or at least, semiautomated literature analysis. For instance, Moro, Cortez, Rita (2015), employed TM to identify relevant terms and topics of business intelligence research applied to the banking industry. Nunez-Mir, Iannone, Pijanowski, Kong, Fei (2016) employed TM methods to demonstrate how automated content analysis could be helpful in synthesizing knowledge from the enormous volume of ecology and evolutionary biology literature. Guerreiro, Rita, Trigueiros (2016) employed TM to study research in cause-related marketing. Park, Nagy (2018) employed TM to study thermal comfort and building control research. Haneem, Kama, Ali, Selamat (2017) employed Data Analytics and TM in a literature review on the topic of Master Data Management to show how those algorithms could assist the process of literature analysis.

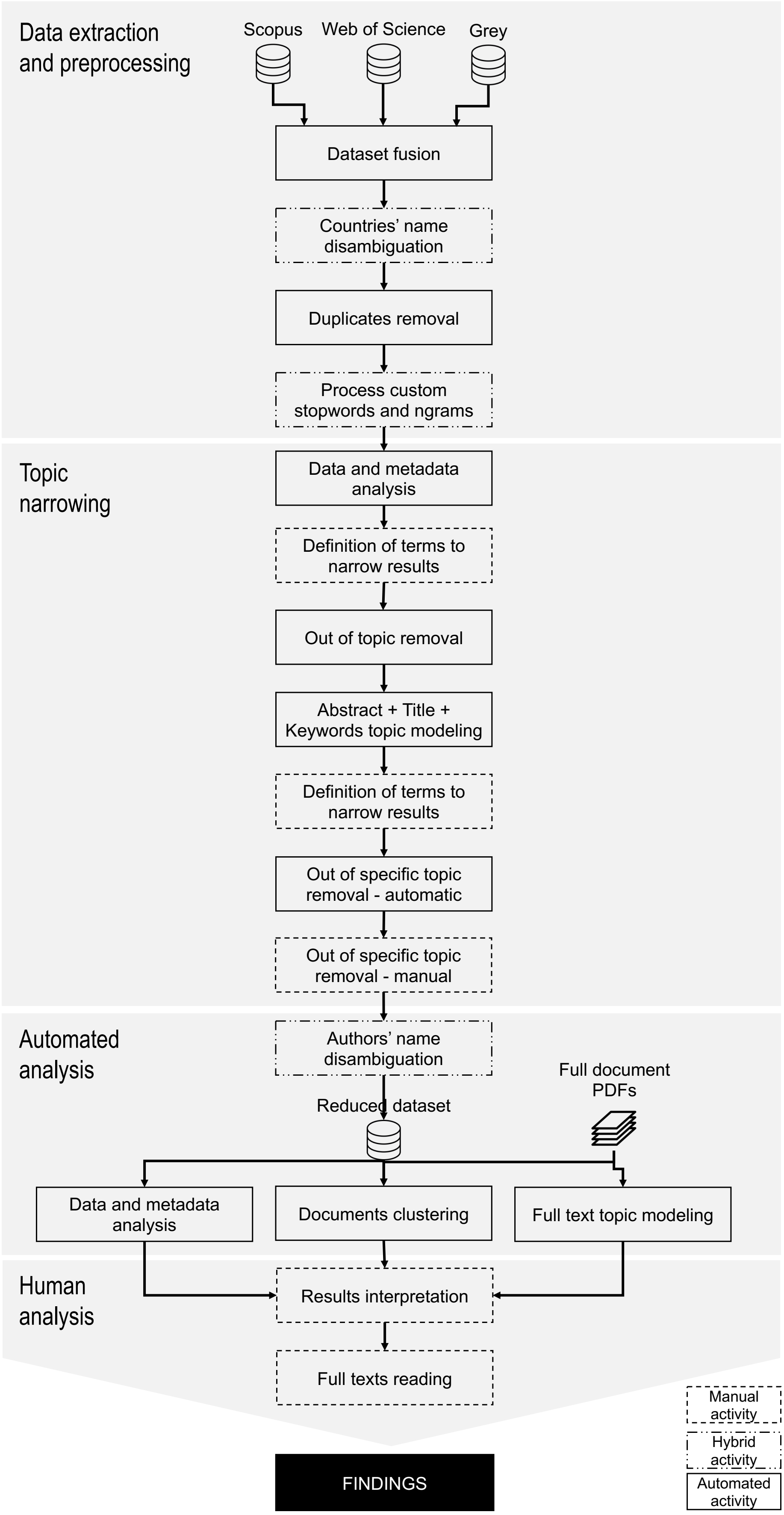
All the above-cited examples of automated/semiautomated literature reviews share the fact that all employed TM. TM seeks to extract useful information from documents collections through the identification and exploration of patterns. While Data Mining (DM) assumes data is stored in a structured format, TM data is essentially stored in a non-structured format. For this reason, TM data requires the application of preprocessing operations to identify and extract features representative of natural language documents (Welbers, Van Atteveldt, Benoit 2017). Due to the importance of natural language processing in TM mission, the latter draws on the advances of other computer science disciplines, like Data Science, to achieve its objectives.

2.1.2.2 Methodology

Literature on the automation of literature review propose rather similar processes to conduct literature analysis (Delen, Crossland 2008; Feng, Chiam, Lo 2017; Haneem, Kama, Ali, Selamat 2017; Nunez-Mir, Iannone, Pijanowski, Kong, Fei 2016; Tsafnat, Glasziou, Choong, Dunn, Galgani, Coiera 2014). However, those processes differ on the algorithms or approaches employed. This proposal is based on the process used by Haneem, Kama, Ali, Selamat (2017), but adapted to the topic under research and whose diagram is depicted in Figure 2. This procedure is divided into four main steps, which in turn are spilt into activities. Some are fully automated activities, others are manual activities, and others are hybrid, i.e., partially automated. The details of these activities are presented in the following sub-sections. In order to explain the rationale behind the methodological choices, some results are presented interleaved with the methods.

All the steps of the experimental procedure presented in Figure 2 were conducted using R (R Core Team 2016), which is a powerful statistical tool with numerous packages developed by its user community to extend its capabilities, designed to facilitate data analysis.

Figure 2 - Procedure workflow diagram


Data extraction and preprocessing
Scopus
Web of Science
Grey
Dataset fusion
Countries' name disambiguation
Duplicates removal
Process custom stopwords and ngrams
Topic narrowing
Data and metadata analysis
Definition of terms to narrow results
Out of topic removal
Abstract + Title + Keywords topic modeling
Definition of terms to narrow results
Out of specific topic removal - automatic
Out of specific topic removal - manual
Automated analysis
Authors' name disambiguation
Reduced dataset
Full document PDFs
Data and metadata analysis
Documents clustering
Full text topic modeling
Human analysis
Results interpretation
Full texts reading
FINDINGS
Manual activity
Hybrid activity
Automated activity

2.1.2.2.1 Data extraction and preprocessing

Quality literature analysis must cover relevant literature on the research topic and should not be confined to one specific research methodology, one set of journals or one geographic region (Webster, Watson 2002). The search strategy is an important component of the analysis of literature. The present approach is based on what Ali & Usman (2018) call an "automated-search", that is, a search strategy that relies on electronic databases keyword searches. The number of databases used, and its type, are essential guidelines to guarantee the quality of the review (Ali, Usman 2018). Thus, guarantee the quality of the review, two well-known databases were employed: Scopus and WoS. These databases cover the majority of sources related to relevant tourism and travel industries research. An adequate selection of keywords must be used in the databases search and the correct construction of the search string (Ali, Usman 2018; Delen, Crossland 2008). Taking in consideration the problems that might arise from the differences among database search engines (Tsafnat, Glasziou, Choong, Dunn, Galgani, Coiera 2014), a simple query was executed on both databases, and the results were filtered to narrow the search to the objective of the research. The search string finds simultaneous usage of the words "booking" and "cancellation" in the title or keywords of publications. In the case of the Scopus database, the words were also searched in the abstract. WoS did not have this functionality. Sometimes the word "reservation" is employed instead of "booking" so it was also searched. Variations in plural and in UK and American English of the word "cancellation" were accounted for. The application of TM for multiple languages presents methodological difficulties. However, given that the most relevant research is published in English, the search was limited to English publications. Since each database has its document classification categories, the type of publications selected in each of the databases was different. For Scopus, the chosen publications' types were: article, article in press, book, book chapter, conference paper or review. For WoS, the chosen publications were of types: article, book, book chapter, proceedings paper or review. The full search strings are shown in Figure 3 and Figure 4, respectively. The term "no-show" was not included in the search string in order avoid the identification of publications that solely address the problem of "no-shows" forecast (quite common in the airline industry as previously introduced). This exclusion of the term "no-show" does not mean no-show models are not pertinent for this research, on the contrary. However, as described in the scope of this dissertation, there is no interest in distinguishing no-shows from cancellations. In this type of static models, adjustment to values is made by resolving the model periodically to account for changes in probabilities over time (Talluri, Van Ryzin 2005).

Scopus search results were exported to a CSV (Comma Separated Values) file using Scopus export functionality. WoS search results were exported to a TSV (Tab Separated Values) file. To assess the validity of each search results, a randomly selected number of publications were checked to ascertain their inclusion based on the search words. Next, the inclusion of known publications on the topic in the search results was also checked.

Figure 3 - Scopus search string

```
TITLE-ABS-KEY ( ( "booking" OR "bookings" OR "reservation" OR "reservations" ) AND ( "cancellation" OR "cancellations" OR "cancelation" OR "cancelations" ) ) AND ( DOCTYPE ( ar ) OR DOCTYPE ( ip ) OR DOCTYPE ( bk ) OR DOCTYPE ( ch ) OR DOCTYPE ( cp ) OR DOCTYPE ( re ) ) AND PUBYEAR > 1989 AND ( LIMIT-TO ( LANGUAGE , "English " ) )
```

Figure 4 - WoS search string

```
(((TS=(("booking" OR "bookings" OR "reservation" OR "reservations") AND ("cancellation" OR "cancellations" OR "cancelation" OR "cancelations")) OR TI=(("booking" OR "bookings" OR "reservation" OR "reservations") AND ("cancellation" OR "cancellations" OR "cancelation" OR "cancelations")))))AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Book OR Book Chapter OR Proceedings Paper OR Review)
```

Timespan: 1990-2018. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC.

One other dataset was manually created, a dataset with studies not usually found in literature databases – the so-called “grey literature”. This type of literature, often neglected, such as dissertations, articles published in obscure journals, online journals, industry journals, or in the different types of internet websites, should also be collected. Otherwise, the literature review should not be considered systematic (Eysenbach, Tuische, Diepgen 2001; Tsafnat, Glasziou, Choong, Dunn, Galgani, Coiera 2014). With this in consideration, this manual dataset was created with a structure equal to the format used by Scopus and stored in a CSV file. The selection of publications to include this dataset was made in two steps. A first step based on a search conducted on google.com using the search strings “booking cancellation prediction” and “booking cancellation forecasting”. A second step, through “snowballing” the references of the publications present in the Scopus and WoS datasets. “Snowballing” is a method that allows the identification of relevant papers by tracking citations of other paper (Greenhalgh, Peacock 2005).

The fusion of multiple databases from different origins is always a challenge due to the differences of the databases structures, data formats and data quality. In light of this and in the scope of this review, a single dataset was created based on the fields described in Figure 5. The creation of this dataset involved a normalization process: the conversion of all text into lowercase - a transformation of all words to a more uniform form (Welbers, Van Atteveldt, Benoit 2017). All the text preprocessing was performed using the “NLP” (Hornik 2017) and “tm” (Feinerer, Hornik 2017) R packages.

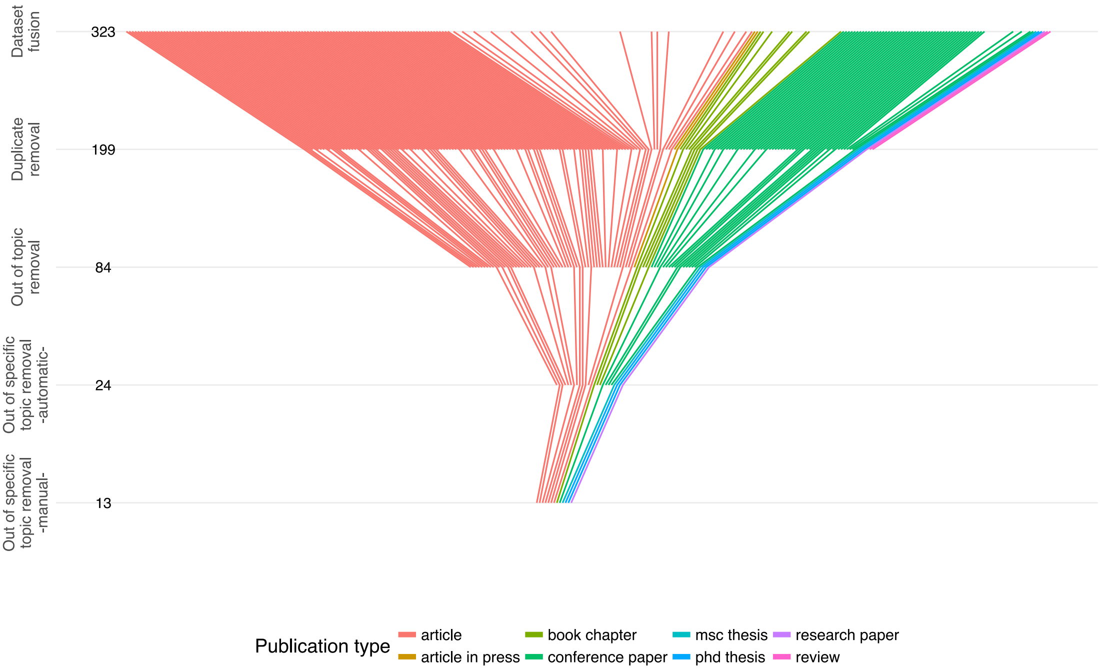
Figure 5 – Datasets' fields match diagram

Scopus/Manual	Fused	WoS
Affilliations	Affilliations	C1
Authors	Authors	AU
Index.keywords	IndexKeywords	
Author.keywords	AuthorKeywords	DE
Document.type	DocumentType	PT
DOI	DOI	DI
Cited.by	Cited.by	Z9
Abstract	Abstract	AB
Title	Title	TI
Abbreviated.Source.title	SourceAbreviation	DS2
Source.title	Source	SO
Year	Year	PY
EID	ID	UT

As it can be observed in Figure 6, the fusion of the three datasets resulted in a total of 323 publications (167 from the Scopus, 149 from the WoS, and seven from the grey literature dataset) of which a substantial part were duplicates. The only common field identifier in all databases is the DOI. However, not all publications have a DOI. Thus, the removal of duplicates was achieved after the titles of the publications were preprocessed, by comparing the titles and then comparing the DOIs. Preprocessing is a process that tokenizes full texts to smaller and specific features, including normalization of words, for improved analysis and enhanced computational performance. Preprocessing text involves the removal of punctuation, removal of numbers, removal of stopwords and stemming. Stopwords are words that are common in a language, e.g., "the" or "a". Stemming normalizes words with different morphological variations, such as verbs conjugation suffixes or the plural of a noun (Welbers, Van Atteveldt, Benoit 2017). Title preprocessing allowed for the capture of duplicate publications that a simple comparison would not. For example, the title of an article in Scopus did not include the initial "The". One exception to the automatic identification of duplicates was found: two publications share the same title and abstract, although having different sources, DOI and authors. A manual verification showed that it was the same publication but presented in two different conferences. The removal of duplicates reduced the dataset to 199 publications (Figure 6).

With the help of a Document-term matrix (DTM), the frequency of common terms was verified. A DTM, or corpus, is a common form for representing a collection of documents, which assigns documents to the rows of a matrix and terms contained in the documents to columns. The cells of the matrix indicate the frequency of the terms in the documents, allowing for the matrix to be analyzed with vector and matrix algebra (Welbers, Van Atteveldt, Benoit 2017). This analysis was necessary to identify terms that, although presenting a high frequency, were not relevant (e.g. "elsevier bv", "all rights reserved", "et al", among others). On the other hand, terms that were relevant but composed of multiple words and should be converted to one-word terms (e.g. "revenue management" or "no-show"). In TM, terms can also be called "n-grams", where "n" indicates the number of words. Single word terms are called "unigrams". Sequences of two words terms are called “bigrams” and so on (Welbers, Van Atteveldt, Benoit 2017). The terms identified as not being relevant were simply removed from the title, author keywords, index keywords and abstract, while the relevant bigrams were converted to unigrams (e.g., “revenue management” was converted to “revman” and “no-show” to “noshow”).

Figure 6 - Publication selection funnel

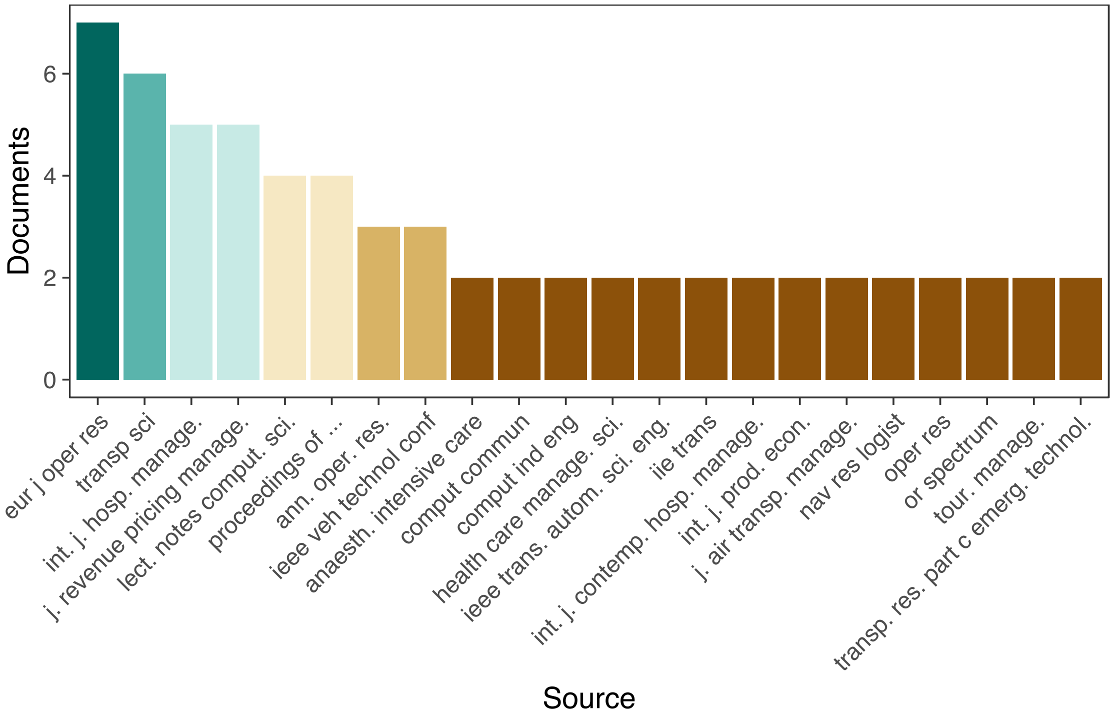


2.1.2.2.2 Topic narrowing

The 199 publications that resulted from the previous step were composed by 129 (65%) articles, 54 (27%) conference papers, 8 (4%) book chapters, 2 (1%) articles in press, , 2 (1%) reviews, 2 (1%) PhD dissertation, and 1 (1%) MsC dissertation. Publications came from 156 different sources. From those sources, only 22 had more than one publication. Considering the sources’ names depicted in Figure 7 it is possible to verify that the publications are from different areas, namely: operations research, hospitality management, medicine and health, transportation and logistics management, among others.

Although all previous studies on the topic of bookings cancellation forecast/prediction could potentially be relevant for the objective of this research, studies from industries not related to the travel industry, or studies from industries who not share some of the hotel industry characteristics, such as variable and uncertain demand, or perishable inventory, would hardly be applied to the hotel industry. Consequently, these non-travel industries related publications needed to be removed from the publications dataset.

Figure 7 - Publications per source (sources with more than one publication)



Two Term-Document Matrixes (TDM)[3] were created using a preprocessed version (removal of numbers, stopwords and stemmed) of all of the abstracts in the dataset to identify the terms to be used on the search for publications related to travel industries. One TDM for unigrams and another for bigrams allowing for the counting of the frequency of each term in the corpus. The R package "wordcloud" (Fellows 2014) was employed to elaborate word clouds to enable the analysis of the terms with frequency equal or above ten (Figure 8 and Figure 9). Regarding unigrams, it is possible to verify that the most frequent terms are related to travel industries (e.g., "cancel" or "reserv") or are general terms (e.g., "model" or "system"). However, some terms like "patient" or "appoint" point to medicine and health industries. A similar pattern seems to emerge in bigrams, where the more frequent terms were related to the travel industries (e.g., "book limit" or "cancel noshow"), but others seem to be related to medicine and health (e.g., "miss appoint") or electronics (e.g., "papr reduct" or "multiplex ofdm").

Considering the analysis of the terms frequency to restrict publications to be searched for, the chosen filter terms are tourism, hotel, airline, aviation, restaurant, golf, railway and "revman". A stemmed version of these terms was then looked for in preprocessed versions of the title, author keywords, index keywords and abstract of all 199 publications in the dataset. Every publication that did not contain one of the terms was excluded from the fused dataset, and that was not part

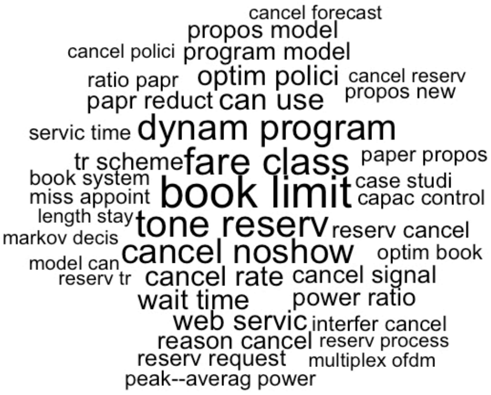
[3] A TDM is a transpose of a DTM, i.e. a matrix where terms are assigned to rows and documents assigned to columns.

of the grey literature dataset, which resulted in filtering the dataset to a total of 84 publications (Figure 6).

Figure 8 - Unigram word cloud

cancel
reserv
model
use
time
system
patient
book
servic
optim
custom
problem
polici
propos
paper
result
can
demand
data
hotel
overbook
number
perform
airlin
process
schedul
approach
method
appoint
cost
develop
provid
rate

Figure 9 - Bigram word cloud

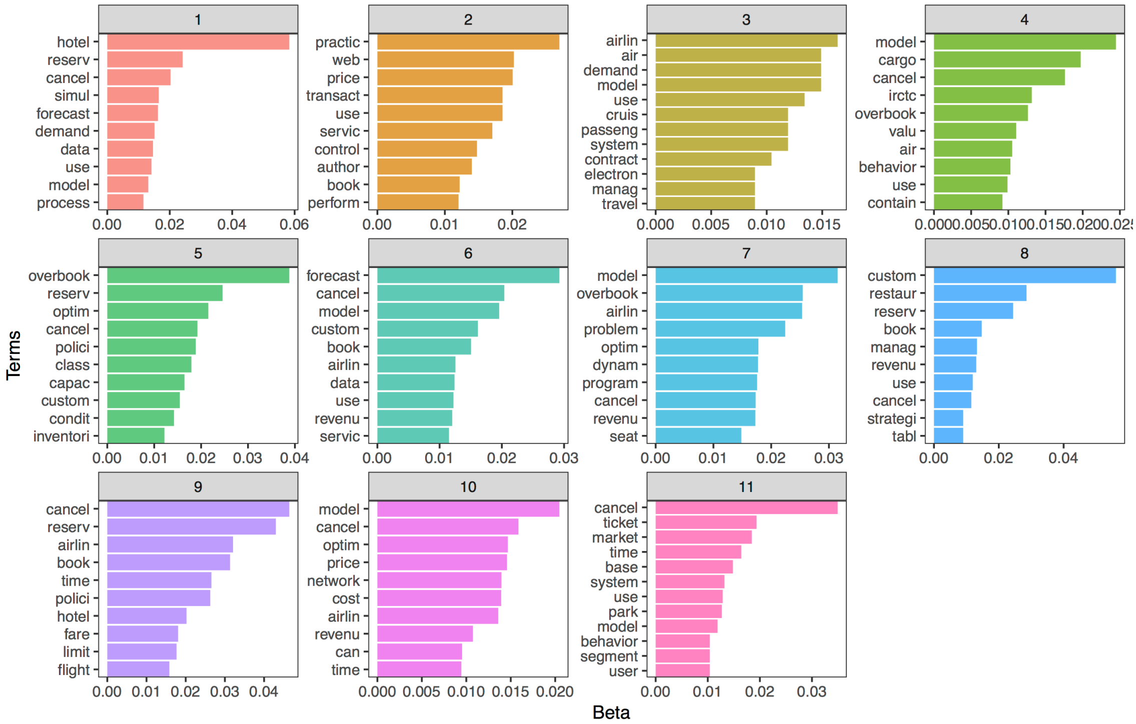


Latent Dirichlet Allocation (LDA) method (Blei, Ng, Jordan 2003) was employed to understand which of the topics were covered by the resulting 84 publications. LDA was applied using the R package “topicmodels” (Grun, Hornik 2011). LDA is the most popular and widely used method for topic modeling (Calheiros, Moro, Rita 2017). LDA is a statistical model that groups text documents based on a classification given by computed measures representing the document’s distance from a given topic and from the document to the known *a priori* (Arun, Suresh, Madhavan, Murthy 2010). The definition of the ideal number of topics was done with the R package “ldatuning” (Nikita 2016). This package uses four different methods to help on the decision of the number of topics. Based the results, the more adequate number of topics was defined to be 11. LDA was then applied to a corpus formed with the preprocessed versions of the title, author keywords, index keywords and abstract of each publication.

Following an analysis of the top 10 terms identified in each of the 11 topics by the LDA beta, the Dirichlet prior on the per-topic word distribution (Figure 10) shows that publications in topic 6 seem to concentrate publications about forecasting, cancellations, and modeling. Topic 10 also have as two of its top three terms “model” and “cancel”, but “forecast” is replaced by “simul”, which may indicate that publications in topic 10 are more related to simulation than forecasting. On the contrary, publications in topic 2 do not seem at all to be related to forecasting/predicting cancellations, as no term related to cancellation, forecast, or modeling appear in the top 10 of this topic.

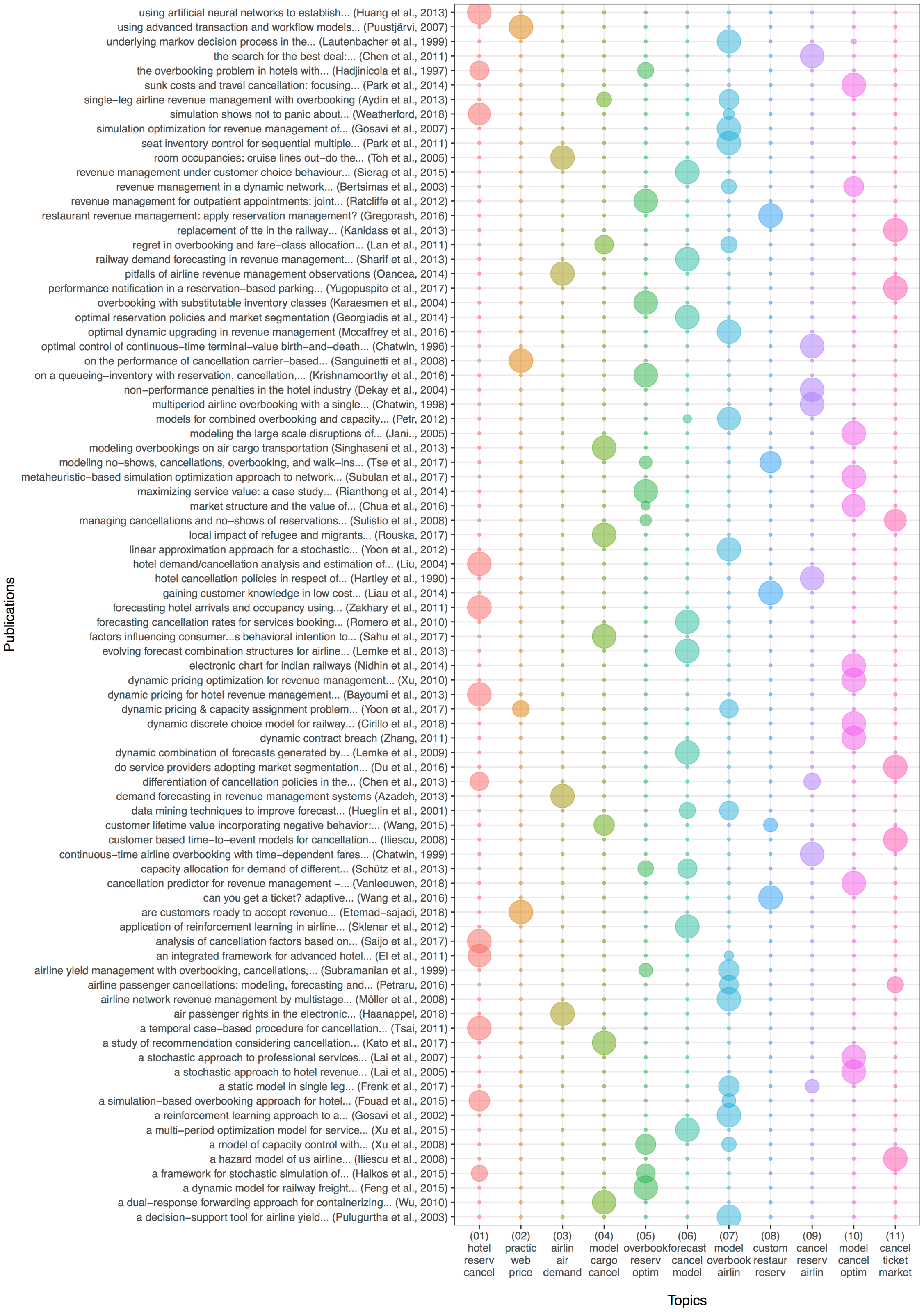
When looking at the probability of a document covering a topic, it is possible to identify which documents cover which topics by using the gamma distribution of LDA (Figure 11**).** An analysis of titles and topics revealed that some publications were not in the forecast/prediction modeling topic. For example, “Local impact of refugee and migrants crisis on Greek tourism industry” (Krasteva 2017) is related to measuring the impact of refugees on tourism (including cancellations, but not modeling them). Based on this, it was decided to narrow the search even further. The new filtering was carried out through a new automated search of terms. This time searching which publications had in their preprocessed title, author keywords, index keywords or abstract, a stemmed version of the term's “prediction”, “forecast”, or “estimation”. The application of this filter resulted in a total of 24 publications (Figure 6).

Figure 10 - Top 10 terms per topic



The reading of the abstracts of the 24 publications hinted that some of the publications were not specific to the development of bookings forecast/prediction models. The reading of the full texts confirmed that 11 of these publications that effectively did not explicitly address this topic. For example, the publication “A decision-support tool for airline yield management using genetic algorithms” (Pulugurtha, Nambisan 2003) is about forecasting the number of seats to allocate to each fare class in airlines. Like this, other publications also took cancellations in consideration, but as a parameter for forecasting overall demand (Liu 2004; Sierag, Koole, van der Mei, van der Rest, Zwart 2015; Zakhary, Atiya, El-Shishiny, Gayar 2011), or were about other subjects, like: the development of revenue management frameworks (Gayar, Saleh, Atiya, El-Shishiny, Zakhary, Habib 2011), exclusively forecasting no-shows (Hueglin, Vannotti 2001; Lan, Ball, Karaesmen 2011), simulators for overall demand forecasting (Fouad, Atiya, Saleh, Bayoumi 2014; Halkos, Tsilika 2015), calculating customer lifetime value (Wang 2015), or were discussing future research in revenue management and current state of the art (Oancea 2014). Consequently, these 11 publications were removed from the dataset, leaving the dataset with 13 publications.

Figure 11 - Publications probabilities per topic



2.1.2.2.3 Automated analysis

To continue with the determination of an adequate publication dataset on bookings cancellation forecast/prediction for travel industries, an automatic verification for disambiguation of author names was carried on. This disambiguation enabled the identification of several authors whose name was written differently in different publications (e.g., only with the first and last names or with the full name). The names were manually corrected for subsequent publication analysis.

The PDF files for all the 13 selected publications were manually downloaded from the publications' publisher website or scientific repositories. A new corpus was then created by including a preprocessed version of the full text of each publication. This preprocessing included several normalization processes, namely: case lowering, removal of numbers, removal of punctuation, removal of the non-informative terms previously identified, conversion of two-word terms to one word (for the terms previously identified) and stemming.

The new corpus was used to classify publications by clusters and topics. Documents clustering is perhaps the most commonly used analysis technique in TM applications (Delen, Crossland 2008). One of the challenges in clustering is determining the number of clusters to be discovered (Kassambara 2017). The R package "factoextra" (Kassambara, Mundt 2017) was used to identify and analyze clusters. For determining the number k of expected clusters for the dataset the "elbow" and the "average silhouette" methods were used. Albeit using a weighted term frequency-inverse document frequency (also known as tf-idf) DTM, results were k=1 and k=2 for the "elbow" method and for the "average silhouette" methods, respectively. The reason for this is probably associated with the low number of publications, which also explains the result for topic modeling (obtained with the "ldatunning" R package) that determined that number of topics in these 13 publications should be between 8 to 12.

2.1.2.3 Results and discussion

The 13 selected publications were distributed between 6 of the 11 topics identified in Figure 11. Four publications were attributed to topic 6, a topic that precisely had as 3 top terms, terms in the root of this dissertation: forecast, cancel and model. Two documents were also attributed to each of the topics 7 and 10, which also had the term "model" has one of their top 3 terms. However, the other words in topic 7 indicate the topic is related to modeling overbooking in the airline industry. As for topic 10, its top 3 terms indicate is somehow related to optimization. Other two publications were attributed to topics 1 and 11. While topic 1 and has 3 top terms, terms related to hotel bookings cancellation, topic 11 was related to the airline and railway industries. The remain publication was attributed to topic 3, a topic related to forecasting demand in airlines.

Table 2 presents a summary of the identified 13 publications, including the indication of methods, the problem addressed and type of data employed.

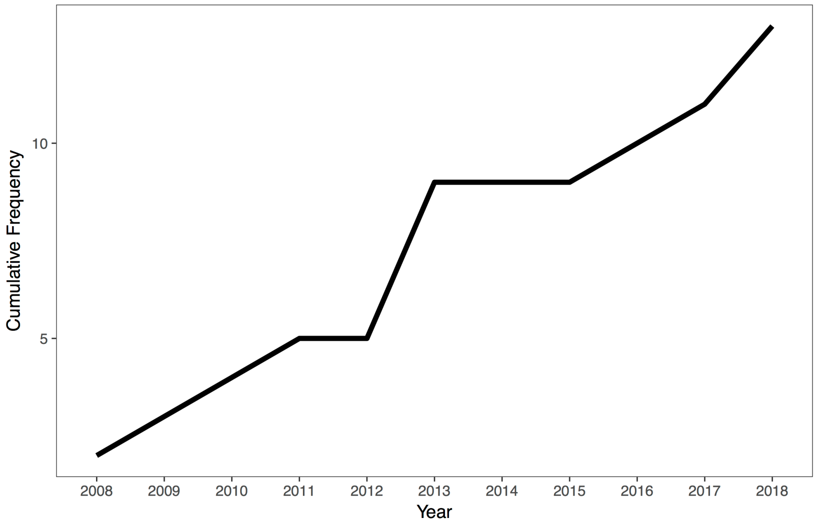
Table 2 - Summary of the 13 final publications (ordered by publication year)

Author (Year)	Methods type	Problem type and algorithms	Data and industries
Iliescu, Garrow, Parker (2008)	Advanced booking	Prediction/classification. Discrete time proportional odds	Ticketing data from Airline Reporting Corporation (ARC). Airline industry
Iliescu (2008)	Advanced booking	Prediction/classification. Discrete time proportional odds	Ticketing data from Airline Reporting Corporation (ARC). Airline industry
Lemke, Riedel, Gabrys (2009)	Advanced booking	Forecasting/regression. Combination of single exponential smoothing, Brown's exponential smoothing and a regression approach	Weekly aggregated booking data from Lufthansa Systems Berlin GmbH. Airline Industry
Morales, Wang (2010)	Advanced booking	Forecasting/classification (for cancellation rate calculation). Average cancellation rate, seasonally averaged rate, logistic regression, C4.4 decision tree, minimum squared expected error tree, random forest, support vector machine and kernel logistic regression	Hotel chain bookings in PNR format. Hotel industry
Tsai (2011)	Combination	Forecasting/regression. Combination of different statistic algorithms	Aggregated railway booking data. Railway industry
Lemke, Riedel, Gabrys (2013)	Advanced booking	Forecasting/regression. Combination of different statistic algorithms and genetic algorithms	Weekly aggregated booking data from Lufthansa Systems Berlin GmbH. Airline Industry
Azadeh, Labib, Savard (2013)	Historical	Forecasting/classification (for cancellation rate calculation). Multi-layer perceptron neural network.	Historical aggregated data of railway operator. Railway industry
Azadeh (2013)	Historical	Forecasting/classification (for cancellation rate cancellation). Multi-layer perceptron neural network	Historical aggregated data of railway operator. Railway industry
Huang, Chang, Ho (2013)	Advanced booking	Forecasting/classification. Back propagation neural network and general regression neural network	Restaurant booking data from a western chain in Taiwan. Restaurant industry

Author (Year)	Methods type	Problem type and algorithms	Data and industries
Petraru (2016)	Historical	Forecasting and prediction/regression and classification. Five different time series algorithms	Airline simulated data. Airline industry.
Tse, Poon (2017)	Historical	Forecasting/regression. Maximum-likelihood estimation	Daily aggregated booking data from restaurant. Restaurant industry
Cirillo, Bastin, Hetrakul (2018)	Advanced booking	Forecasting/classification. Dynamic discrete choice model	Intercity detailed ticket railway data. Railway industry
van Leeuwen (2018)	Advanced booking	Prediction/classification. Naïve Bayes, logistic regression, decision tree and random forest	International hotel chain detailed reservation data for 7 hotels. Hotel industry

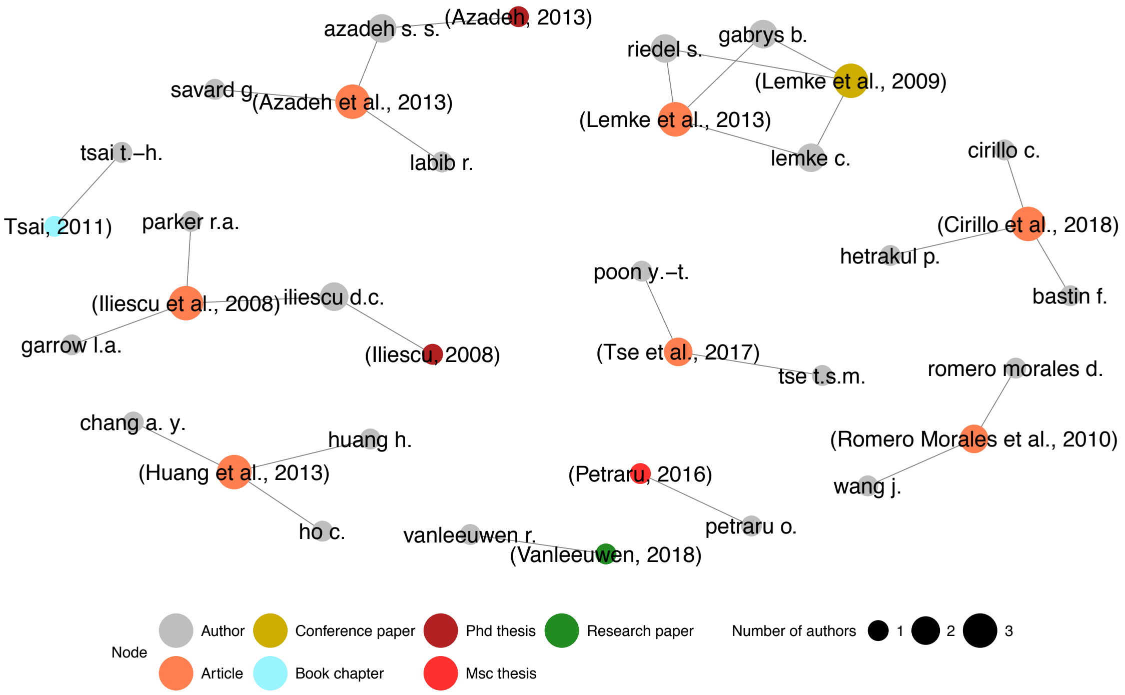
As it can be seen in Figure 12, the first documents specific to bookings cancellation forecasting/prediction modeling where only published in 2008. Since then, this number have been increasing steadily, with the exception of 2012, 2014 and 2015. Understandably, as for the topic of overall demand forecast modeling, for the particular topic of cancellation forecast/prediction modeling, the airline industry is the industry on which more publications focus. From the 13 publications, 5 used airline data, 4 railway data, 2 restaurant data, and 2 hotel data.

Figure 12 – Publications published over the years



An authors' network diagram shows the sparsity of research and the diversity of the type of publications on the subject of bookings cancellation forecasting/prediction (Figure 13). The figure shows research on the topic is confined to a few groups of authors, with no collaborations between them. It is also possible to verify that some authors have more than publication on the subject, being one of the publications, in two cases, PhD dissertation (Azadeh 2013; Iliescu 2008).

Figure 13 - Authors' network



The network of keywords, as seen in Figure 14, also has a high level of sparsity, with some groups of keywords being employed just in one of the publications, thus only relating between themselves and not with other keywords. The exception is the keywords “revenue management”, “forecasting”, and “cancellation”. This exception suggests that research’s topics diverge by groups of authors.

The cluster analysis, as illustrated in Figure 15, shows a differentiation between a dissertation (PhD and MsC) and other types of publications. While the dissertation from (Azadeh 2013) was included alone in one of the clusters, the other two dissertations (Iliescu 2008; Petraru 2016) were the publications at more distance from other publications on the other cluster. This distinction between dissertation and other publications might be explained by the resemblance of dissertation structures and the difference in size between dissertation and other publications.

As seen in Table 2, almost half of these publications (six in total) employed detailed booking or ticket data. This increasing tendency to employ detailed booking data in forecasting models, in particular of data in the PNR format, in detriment of time series aggregated data is related to the advances in technology and forecasting algorithms (Morales, Wang 2010; Petraru 2016). Some publications employ data in the ARC format instead of PNR format. PNR and ARC formats are both standards from the airline industry, with PNR being widely used in demand forecasting. The main reason could be its origin. While ARC data is based on tickets issued, PNR data is based on bookings made. The first is triggered by financial events (e.g., purchases, refunds and exchanges), while the second is triggered by reservation systems (e.g., bookings and cancellations) (Iliescu 2008). ARC data, as reported by Iliescu (2008), includes 21 fields: carrier, issue date, departure dates (inbound and outbound), new departure dates (inbound and outbound according to exchange event), exchange event, refund date, void date, exchange fee, fare, fare different, new flight number (in case of an exchange), ticketing class new In case of exchange), ticketing class code (2 codes), ticketing class code new (2 codes in case of exchange), type of trip (one way or round trip). On the other hand, PNR data, although also specifically built for the airline industry, does not have a not so rigid format. Operators can include their own fields, according to the detail they want. However, operators have to comply with the guidelines on what information should be included in PNR fields. While passenger identification, flight details, meal preferences, health issues should be present, information not relating to the travel, such as ethnic origin, political opinions, religious beliefs, marital status, should not be present. Other fields that usually are included in PNR records is baggage information, check-in information, go-show information, no-show information, number of passengers, frequent flyer number and status, travel agent details (International Civil Aviation Organization 2010).

Figure 14 - Keywords' network

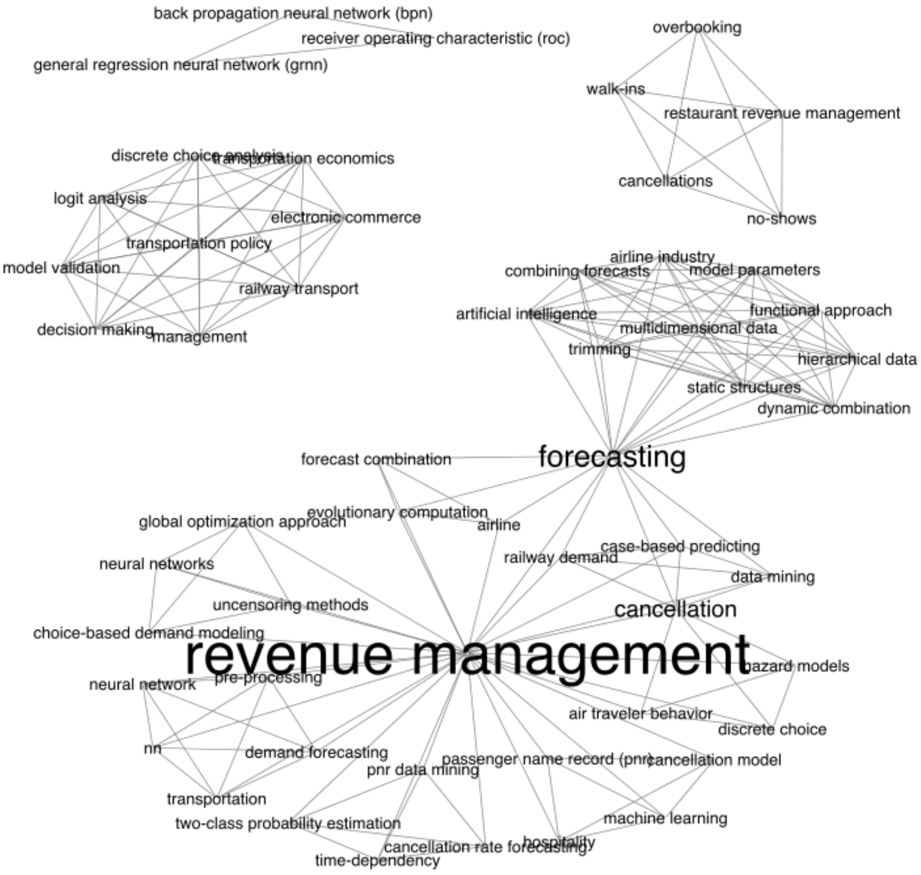
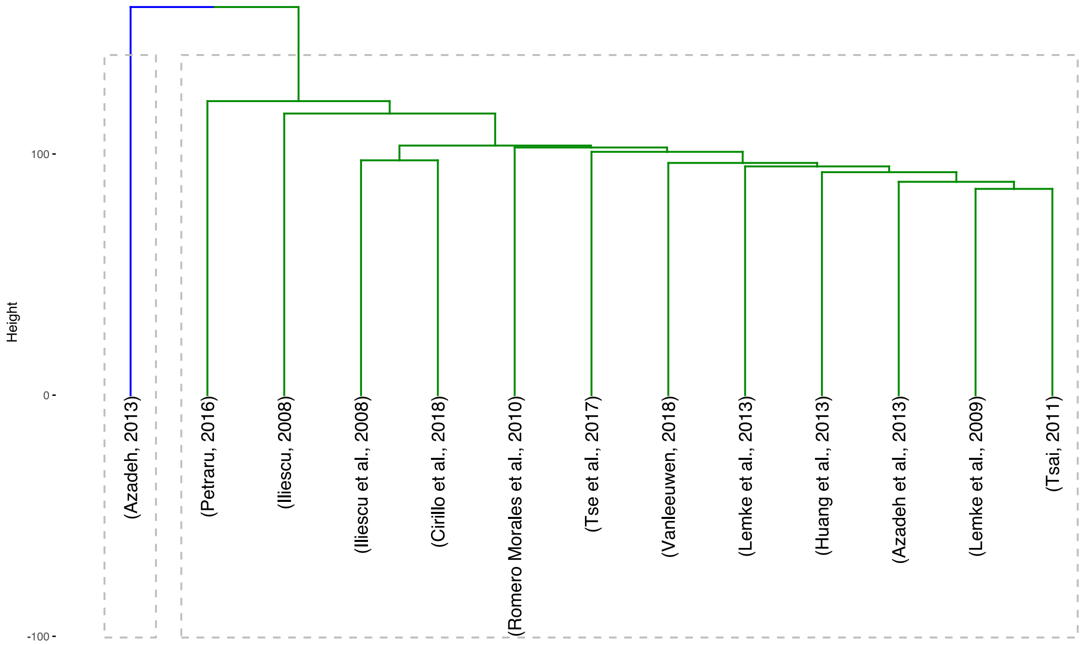


Figure 15 - Publications' clustering



Costs associated with the storage and processing of detailed booking data, as data in the PNR format, has now been mitigated by the development of technology in recent years (Petraru 2016; Tsai 2011). Therefore, the use of detailed booking data instead of aggregated times series historical data not only has the power to improve the accuracy of forecasts (Hueglin, Vannotti 2001; Petraru 2016) but also has the power to allow the building of classification prediction models. In turn, cancellation prediction models, in addition, to allowing the classification of the cancellation outcome of each booking, also allow the understanding of each feature in the data influences cancellations, i.e., allow the understanding of cancellation drivers (Morales, Wang 2010; Petraru 2016). From the thirteen publications identified in Table 2, six employed classification algorithms, however, only four used classification algorithms to understand cancellation drivers, i.e., understanding the past – a prediction problem (Iliescu 2008; Iliescu, Garrow, Parker 2008; Petraru 2016; van Leeuwen 2018). Huang, Chang, Ho (2013) treated the problem as a classification problem as well but did not pursue the identification of cancellation drivers. The remaining two publications who employed classification algorithms used these algorithms to forecast cancellation rates and cancellation deadlines, that is, treated the problem as a forecasting/regression problem and not as a classification problem (Cirillo, Bastin, Hetrakul 2018; Morales, Wang 2010). The reason for this could be the authors' belief *"that it is hard to imagine that one can predict whether a booking will be canceled or not with high accuracy simply by looking at PNR information"* (Morales, Wang 2010, p. 556). Still, the results of Huang, Chang, Ho (2013), and van Leeuwen (2018) contradicted this. Huang, Chang, Ho (2013) back-propagation neural network model for predicting cancellations in restaurants achieved 0.809 in

AUC, 0.751 in *Accuracy* and 0.389 in *Precision*[4]. Using Random forest algorithms for predicting hotel bookings cancellations, van Leeuwen (2018) achieved *Accuracy* values ranging from 0.778 to 0.890, and *Precision* values from 0.823 to 0.899. The higher results obtained by van Leeuwen (2018) are probably explained by the effort put into feature selection and feature engineering. Huang, Chang, Ho (2013) employed 12 features from customer and spend attributes, namely year, month, day, whether the day was a holiday, gender, age, income, education level, marital status, place of residence, cancellations record, and the cumulative number of cancellations. However van Leeuwen (2018) employed a much more detailed dataset, with 23 features, which included, for instance, information on room rate, rate plan, meal plan, distribution channel, type of booking (group or transient), but also employed dataset fields to engineer other variables, like the email address to identify repeating guests. In fact, van Leeuwen (2018) seems to base his study in some of the concepts developed in one of the first publications that resulted from this dissertation, which is one of the references of the author's study (Antonio, Almeida, Nunes 2017).

These two publications (Huang, Chang, Ho 2013; van Leeuwen 2018) were also the only two from the thirteen publications who combined the use of detailed booking data with advanced classification algorithms, which is a strategy that can be used to implement bottom-up forecasts/predictions. For instance, for the booking prediction cancellation problem, one only prediction model can generate not only each booking outcome prediction but also aggregated predictions. By adding up the outcome of bookings predictions per distribution channel, segment, or other aggregation levels, it is possible to have predictions at intermediary levels and global level. However, none of the publications explored or addressed the possibility of using each booking cancellation outcome prediction to calculate net demand at different aggregation levels.

2.2 Overall discussion and objectives

Despite the recognized importance of bookings cancellation forecast/prediction models to forecast demand, the preceding section confirmed what Chen (2016) reported: that so far, only a few studies have tested or developed cancellation models, particularly for the hotel industry.

Although revenue management literature recommends the use of advanced scientific methods in forecasting/prediction problems, there are not many examples of their use yet. In overall demand forecasting, the exception is the use of neural networks (Law 2000; Talluri, Van Ryzin 2005; Weatherford, Gentry, Wilamowski 2003; Zakhary, Gayar, Ahmed 2010). In bookings cancellation forecasting/prediction modeling, as seen in Table 2, neural networks are also used in three publications (Azadeh 2013; Azadeh, Labib, Savard 2013; Huang, Chang, Ho 2013). Other algorithms like decision trees, random forest, or support vector machine are only used in two publications (Morales, Wang 2010; van Leeuwen 2018). These show that only 5 out of 13

[4] For those not so familiarized with machine learning metrics employed in supervised classification problems a brief description of these metrics is presented in Appendix A.

identified publications on the subject of bookings cancellation forecasting/prediction modeling employed advanced machine learning algorithms.

Given PNR's format flexibility and extendibility, its popularity is comprehensible. Because the PNR format was designed for airlines, it does not include important hotel information. For example, for hotels, flight details or type of trip fields, should be replaced by other fields such as departure date, room type reserved, room type occupied, details of the age of persons/babies, detailed loyalty information (e.g., previous cancellations or no-shows), distribution channel, type of booking (group, transient, or party), segment information, among others. In order to improve their performance, models could also employ data from other sources (McGuire 2017; Pan, Yang 2017a). Variables that represent the business problem correctly can reduce the need for modeling specialization and extensive experimentation, thus obtaining better results (Abbott 2014; Domingos 2012).

This highly detailed data, combined with advanced machine learning algorithms, has the potential to build better cancellation prediction models. Additionally, albeit most high-performance machine learning algorithms are fundamentally a black box that generates highly complex prediction equations (Kuhn, Johnson 2013), some algorithms' outputs are of easier understanding for humans (Abbott 2014; Hastie, Tibshirani, Friedman 2001; Kuhn, Johnson 2013). The understanding of these algorithms' models outputs allows modelers to comprehend the predictive power of the different models' inputs, i.e., allow the development of prediction models – models that now only allow forecasting, but also comprehension of the past. Understanding cancellation drivers, although being an important issue for the development of better cancellation policies (Chen 2016; Morales, Wang 2010) is also an understudied subject.

The two main research questions of this dissertation will precisely address the problems mentioned above of the scarcity in studies specific for the hotel industry, which combine the use of hotel detailed booking data with data from multiple sources, with advanced machine learning algorithms to build bookings cancellation prediction models:

RQ1. Could a booking's cancellation prediction model that uses PMS data display better results than a model that uses PNR data?

RQ2. Could this model be improved with the inclusion of data from additional sources?

Contrary to what Morales, Wang (2010, p. 556) said that *"in the revenue management context, the classification or even probability of cancellation of an individual booking is not important"*, it is the author theory that the prediction of the cancellation outcome of an individual booking is important. If a hotel identifies a booking which is going to cancel, it could contact the customer to try to prevent the cancellation or even to obtain an early confirmation of the cancellation. Both results would be significant in terms of revenue management. The test of this hypothesis and also the understanding of how bookings cancellation prediction models could be implemented in a real production environment will be the subject of the dissertation third research question:

RQ3. Can such model be integrated into an hotel RMS?

2.3 Summary

Since the focus of RMSes is to help revenue managers make better demand-management decisions based on advanced scientific methods and technologies, rather than based on guesswork and intuition (Garrow, Ferguson 2008; Talluri, van Ryzin, Karaesmen, Vulcano 2008), building better cancellation prediction models can help revenue managers improve their decisions. As presented, the literature on this topic is limited, and most do not employ advanced scientific models and technologies, such as machine learning.

Bookings cancellation prediction models that make use of advanced scientific models and technologies could help revenue managers identify bookings with high likelihood of canceling, which could allow revenue managers to contact those bookings to try to prevent cancellations. At the same time, bookings cancellation prediction models contribute to better demand forecasts (overall and disaggregated by distribution channel, segments, or other levels), which in turn allow better overbooking decisions. The development of these models could also contribute to a better understanding of cancellation drivers, which can be of significant importance in the development of better cancellation policies.

In an effort to improve bookings cancellation prediction models, this dissertation will use not only bookings data but also data from additional sources, like weather forecast, competition prices and rooms availability, among others. Because the collection of data from these additional sources takes time, the development of the models was divided in two phases. A first phase, detailed in Chapter 3, makes use of PMS data from four hotels to develop bookings prediction classification models and assess its performance and limitations. A second phase, detailed in Chapter 4, combines PMS data with data from additional sources to develop improved models, assess models' performance and understand cancellation drivers.

While Chapters 3 and 4 address RQ1 and RQ2, if detailed booking data and data from other sources can be helpful to predict bookings cancellations, Chapter 5 addresses RQ3, how could the models previously built be integrated in an RMS, their performance in a real environment and their impact on revenue management decisions.

3 EXPLORATORY MODELS

As substantiated in the previous chapter, advances in technology and forecasting algorithms, together with the decrease of the costs associated to store and to process large amounts of data, fostered the use of detailed booking data in revenue management forecasting models in detriment of time series aggregated data. However, despite the recognized importance of being able to predict booking's cancellation, only a few studies employ machine learning classification algorithms to build booking's cancellation predictive models, particularly for the hotel industry. This chapter highlights the first efforts to fulfill this gap and, especially, answer RQ1: on "could a booking's cancellation prediction model that uses PMS data display better results than a model that uses PNR data?". Understanding which paths should be pursued to answer other research questions such as: whether the use of data from other sources can contribute to the improvement of models and how models can be deployed, is also important. The development of the exploratory models intended to perceive problems related to data collection, data quality, data preparation, modeling and assess the models' performance. One other important objective was that of identifying existing limitations to be overcome in further work. Ultimately, the development of these exploratory models also had the objective of obtaining results that could be disclosed in scientific publications, seminars, and conferences in order to obtain feedback from other researchers and practitioners.

After an introductory section on the elaboration of the exploratory models, a detailed description of the methods and materials employed is presented. Next, the main results achieved will be discussed and the chapter ends with a summary of the work carried out and its impact in the evolution of the research.

3.1 Introduction

Detailed booking data, extracted from the PMS databases of four resort hotels (located in the resort region of the Algarve, Portugal) was used to build the exploratory models. Data spans from 2013 to 2015. Since all the hotels required anonymity from now on hotels will be designated as H1 to H4.

To model the full cycle of development – from data collection, feature selection, and dataset creation, to model development and evaluation - the well-known process model CRoss-Industry Standard Process for Data Mining (CRISP-DM) (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, Wirth 2000) was employed. CRISP-DM, as SEMMA, another process model, seems to be an implementation of an older process model - KDD - but more complete and better documented (Azevedo, Santos 2008). Probably because of its completeness and open standard nature, CRISP-DM is one of the most-used process models in data mining, data science and predictive analytics projects (Abbott 2014; Piatetsky 2014).

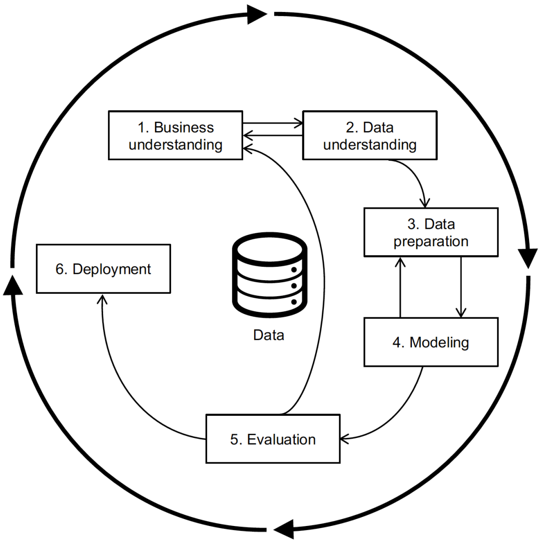
As depicted in Figure 16, the sequence of the six CRISP-DM phases is not rigid and involves going back and forth, with the outcome of one phase indicating which should be the next phase to be performed. The arrows connecting the phases illustrate the most important and frequent dependencies. Until the deployment of a model, multiple iterations between different phases are usually necessary. The outer arrows symbolize the cyclical nature of predictive analytics projects. However, projects do not end when models are deployed. Lessons acquired from modeling the process and its deployment are reincorporated in the model's continuous improvement (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, Wirth 2000).

The six phases that make up the process model are (Figure 16):

- **Business understanding**: an initial phase where project requirements and objectives are studied from a business perspective and converted into an analytics project, resulting in the design of the plan to achieve the objectives;
- **Data understanding**: begins with an initial data collection and continues with activities required to enable modelers to become familiarized with data, including finding patterns, tendencies, and anomalies;
- **Data preparation**: comprises all actives related to the creation of the final dataset (also known as modeling dataset);
- **Modeling**: preparation of the dataset for modeling and application of chosen modeling algorithms, including parameters calibration;
- **Evaluation**: assessment of the models' performance according to the objectives initially set to determine if models have quality to be deployed.

- **Deployment**: application of the model in a real production environment.

Figure 16 - Phases of the CRISP-DM process model



Adapted from Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, Wirth (2000)

The next section offers a detailed description of the execution of each of the phases to construct the exploratory models.

3.2 Process model

Without entering in too much detail, this section offers a comprehensive explanation of the completion of the different phases of the process.

Although the process model phases are sequentially described over the following sections accordingly to the CRISP-DM process model, the outcome of some phases required the flow of work to go back and forth between previous phases.

3.2.1 Business understanding

All four hotels are four and five-star hotels, ranging, in size, from 86 to 180 rooms. All have, at least, one bar and one restaurant. H2 and H3 are mixed-ownership units—besides renting rooms owned by the hotels' management companies, they also rent rooms that were sold in a timeshare or fractional ownership schemes. The summer months, from July to September, are considered as being high season. H1 closes temporarily during the low season, though not regularly. H4 also closed for renovations during a small period.

Cancellation ratios in all hotels have been increasing almost every year, ranging from a minimum of 8.8% to a maximum of 26.4% (Figure 17). These values agree with what was observed by Morales, Wang (2010). Cancellations in these hotels, as portrayed in Figure 18, totalized a value over 6.3 million euros between 2013 and 2015. Understandably, the high cancellation ratios and the amount of revenue lost to cancellations impose high uncertainty for hotel revenue management, substantially influencing pricing and inventory allocation decisions, especially in high-demand dates. As a result, revenue managers need to improve their net demand forecast performance and to gather a better understanding of booking cancellations patterns, tendencies, and anomalies. Considering previous results obtained in the development of bookings' cancellation classification prediction models (Huang, Chang, Ho 2013), it was decided that these models should achieve a prediction *Accuracy* above 0.8 and an AUC also above 0.8 - commonly considered a good prediction result (Zhu, Zeng, Wang 2010).

Figure 17 - Cancellation ratio per year

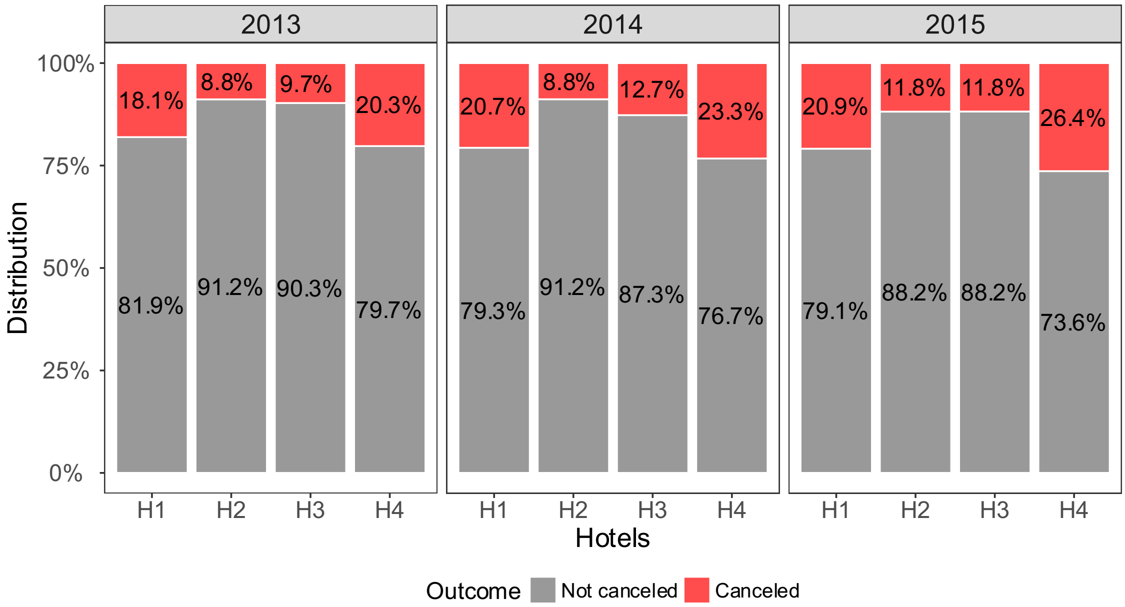
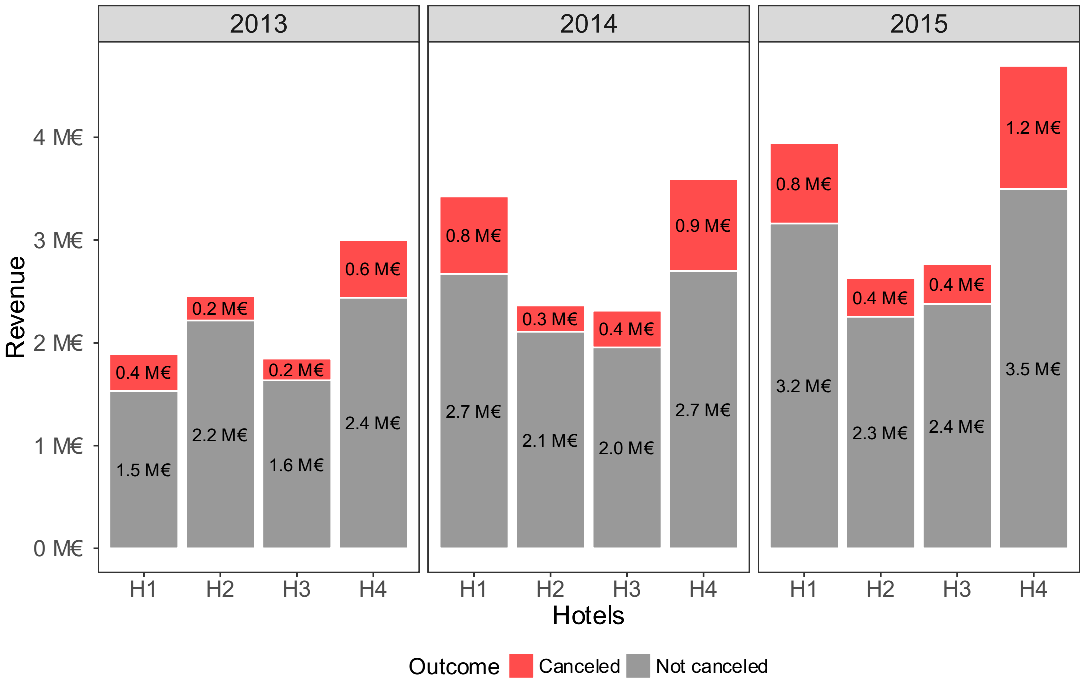


Figure 18 - Revenue per cancellation outcome



A mix of local and cloud resources was employed to build the models. Since all hotels' PMS ran on Microsoft SQL Server databases, copies of these databases were gathered for data extraction. The databases were installed on a MacBookPro computer, with a 2Ghz Intel Core i7, 16Gb of RAM, that ran Mac OS X and Windows 10. Data extraction and transformation was achieved using Structured Query Language (SQL) queries. Data understanding and data preparation phases were conducted using R, chosen due to its high extensibility, which allows it to be a prevalent language for statistical computing and data visualization (R Core Team 2016). Microsoft Azure Machine Learning was used in the modeling and evaluation phases. This platform was selected due to its rich functionality support, including easiness of use, availability of popular machine learning algorithms, powerful model evaluation, and experimentation tools, but mostly because of its capability to make use of cloud computing to deliver fast results, reliably and securely (Barnes 2015).

3.2.2 Data understanding

Since all hotels' PMS are from the same brand, the database structure of the four hotels is very similar. Despite this similarity, specificities of each database and particularities of each hotel operation had to be studied before data extraction SQL queries could be built. The selection of features to include in the datasets was a demanding task. PMS databases data are much richer and more diverse than data in PNR base format, which difficult the selection process. The selection involved a combination of domain knowledge with knowledge from previous studies who identified factors that influenced cancellations, such as the Lead time[5], distribution channel, region of origin of the customer, season for the stay, duration of stay, customer type, or cancellation policy (Chen, Schwartz, Vargas 2011; Liu 2004; McGuire 2017; Morales, Wang 2010; Talluri, Van Ryzin 2005).

A good selection of features facilitates data visualization and data understanding, contributes to reduction of measurement and storage requirements, reduces training and application times, and reduces the risk of falling into the curse of dimensionality – when the amount of data conjugated with a high number of predictor features requires a high computational cost (Abbott 2014; Guyon, Elisseeff 2003). Feature selection, and mainly feature engineering, can contribute positively to the accuracy of prediction models due to the information gain obtained from the association of multiple input variables (Abbott 2014; Guyon, Elisseeff 2003; Kuhn, Johnson 2013). As a matter of fact, feature engineering is considered the key factor in the success of machine learning projects (Domingos 2012). In feature engineering, creativity, intuition, and domain knowledge are as important as technical knowledge.

[5] "Lead time" or "booking window" are terms employed in hotel revenue management to define a measure calculated as the number of days between the date of reservation and the date of service provision (the arrival date in room bookings).

Multiple iterations between the different phases of the process were required to define the final set of features to be included in the datasets. Predictive modeling datasets are usually two-dimensional, comprised with rows and columns, where rows represent the unit of analysis, and columns represent the measure of the feature (Abbott 2014). In this case, the unit of analysis (row) is one booking, and the measure (column) the value of each feature. Although feature engineering is usually performed at the data preparation phase since some manipulations are easier and faster to make at the data collection point, and because feature engineering reduces storing and processing requirements (Guyon, Elisseeff 2003), some features included in the datasets resulted from some sort of data manipulation at their collection. For example, there were no fields in the PMS' databases with the Average Daily Rate (ADR) of each booking. To create a feature with that information was necessary to consult the price table and discounts associated to the booking for each of the nights and then divide its sum by the number of nights. The final SQL data collection queries results were saved to Comma Separated Values (CSV) datasets.

Data visualization and summary statistics are at the core of data understanding. Summary statistics, like the mean, standard deviation (SD), or distribution analysis, can be the simplest way to gain insight into features (Abbott 2014). Summary statistics of each of the hotel's dataset are presented in Appendix B. The statistics were produced with the "skimr" R package (McNamara, Rubia, Zhu, Ellis, Quinn 2018). A detailed description of each feature, including the indication if the feature results from an input variable or if the feature was engineered from one or more input variables is described in Appendix C.

Summary statistics showed that, despite some abnormalities, the overall quality of the data for all the hotels was good. None of the hotels' datasets presented missing values, the observations represented all bookings in the hotels' databases, the levels of categorical features did not present multiple values with the same meaning, and data was properly formatted. For numeric/integer features, the abnormalities were essentially outliers that can be explained by the way hotels work. For example, H1 and H2 presented a negative *ADR* at the percentile 0, and all hotels presented a maximum *ADR* (percentile 100) way above their mean and their 75 percentile. The reason is that, when making corrections or adjustments, including groups or multiple bookings from one travel agency, hotels create fictitious bookings but process the corrections and adjustments in one of the bookings. How the hotels perform the creation of group or small parties' bookings is one other reason. H1 and H4 create bookings with multiple rooms and then, only when they have a confirmation of the guest names, do they transform those bookings into individual bookings. This process, as illustrated in Figure 19, means that most bookings with more than one room, and usually with two or more adults, are just group or small parties' bookings that were canceled. Data also showed that a higher number of children or babies were usually associated to the bookings with more than one room. *AgeAtBooking* summary statistics also show an operational problem. For all the hotels, there are observations pertaining to guests with, supposedly, more than 100 years old when the booking was made. In some cases, even with more than 200 or with a negative number of years of age. Many of the bookings showed 0 (zero) as the age. This result seems to indicate that there are, not only errors in the filling of the birthdate, as well as a leakage problem, i.e., a problem where the value of the feature may be leaking future information. In this case, the birthdate is mostly only filled in non-canceled bookings. This problem is visible in the histograms of *AgeAtBooking* by booking outcome (Figure 20). *CanceledTime* also leaked the booking outcome as it assumed the value of -1 for all non-canceled bookings.

Figure 19 – Booking outcome per rooms quantity and number of adults

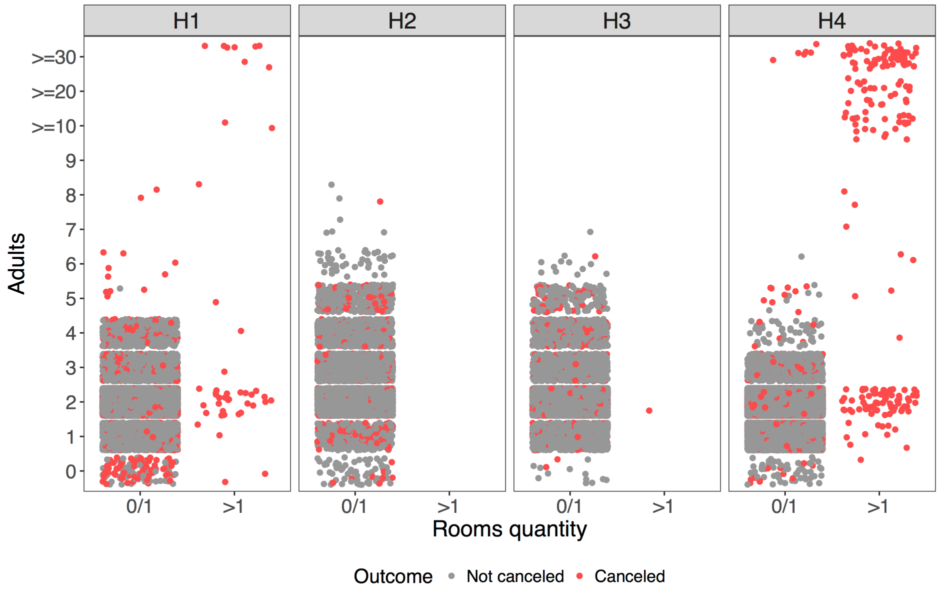
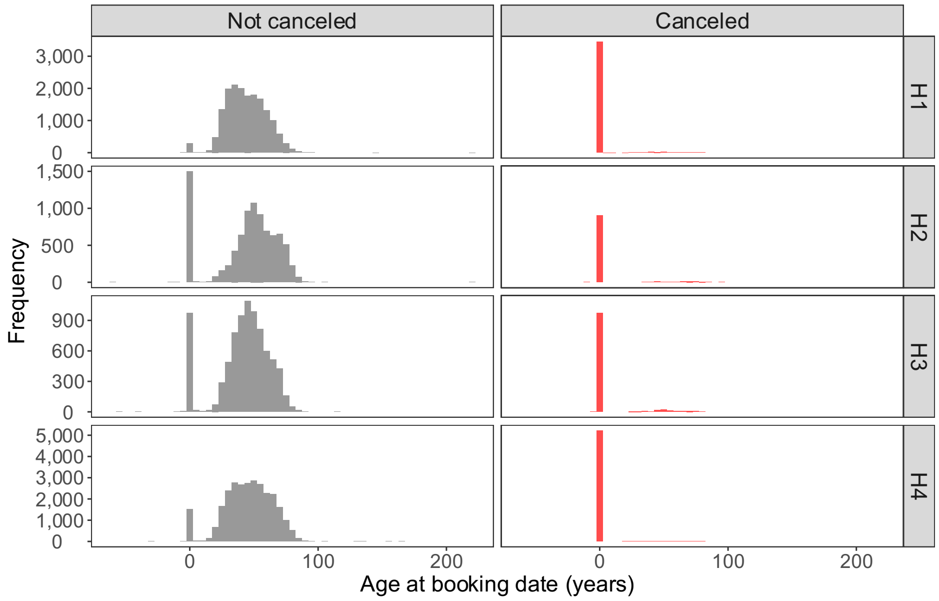


Figure 20 - Distribution of age at booking date



Other examples are just operational outliers, that is, they are real bookings but with some parameters outside the normal range - for example, bookings made a long time in advance (very big *LeadTime*), bookings for an extended period (very big *LengthOfStay*, *StaysInWeekendNights*, and *StaysInWeekNights*), guests that have frequently booked the hotel previously (*PreviousBookingsNotCanceled*, *PreviousCancellations* and *PreviousStays*). However, the visualization of the relationship between these features shows different patterns for all the hotels that are somehow similar. In the case of the relation between *LengthOfStay* and *LeadTime*, as depicted in Figure 21, no particular leakage problem appears. Nevertheless, cancellations seem to increase as *LeadTime* increases. Except for H4 (Figure 22), where there are not many bookings from customers that canceled previous bookings, the number of previous cancellations seems to be a good predictor of cancellation.

Figure 21 - Outcome per length of stay and lead time

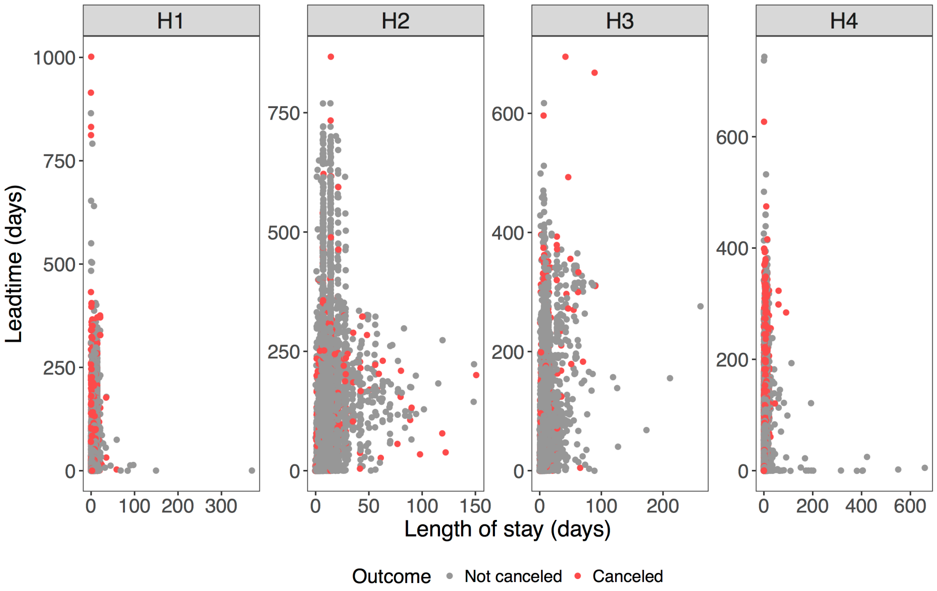
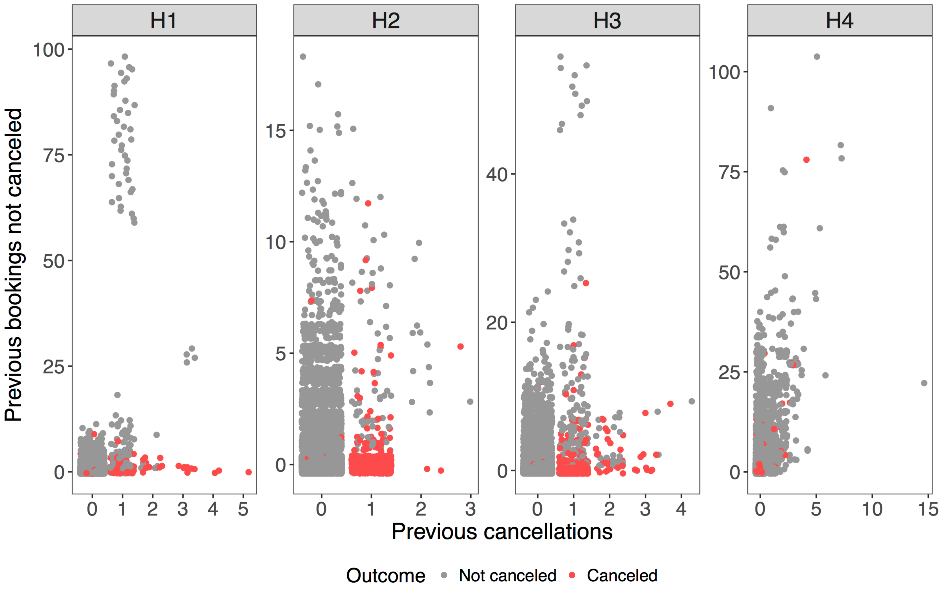
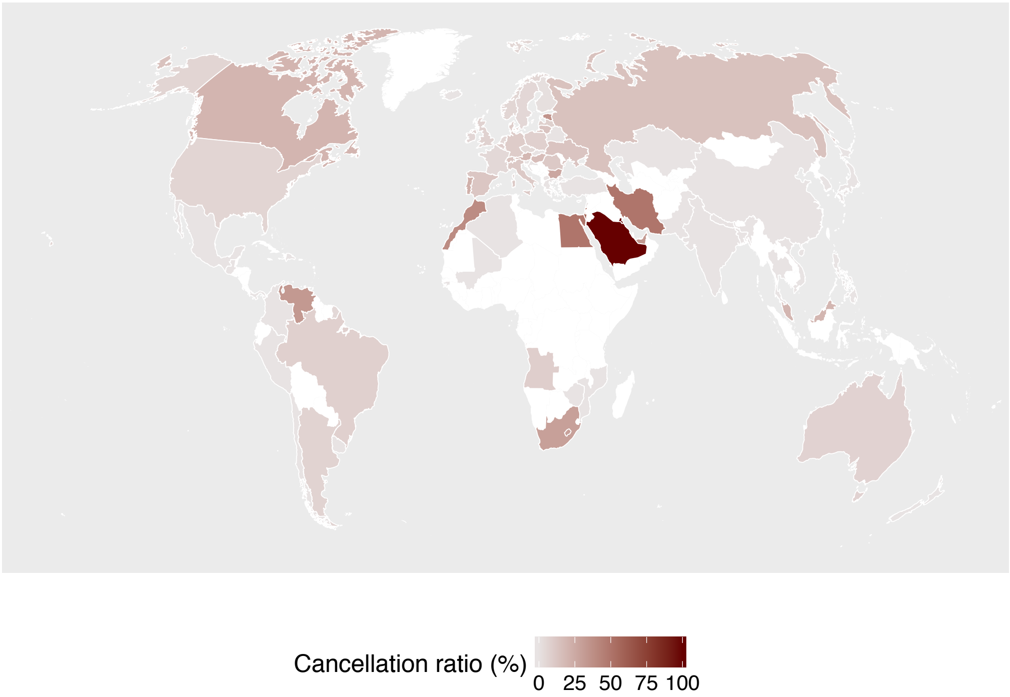


Figure 22 - Outcome per customer prior booking history



Datasets summary statistics of categorical features exposed substantial differences between hotels concerning marketing/segmentation classifications (e.g., distribution channels or market segments) and fundamental features (e.g., agencies, room types, or meal types). These differences served as an indicator that specific models had to be built for each of the hotels, that is, a global model would not fit all the hotels. Other categorical features, like the *Country*, although having the same designations in all hotels, had different patterns in relation to cancellations. For instance, as seen in Figure 23, for H1, the booking cancellation ratio in reservations issuing from some middle eastern countries and some northern African and south American countries is very high. Data also show that, for all the hotels, the major part of the canceled bookings is classified as belonging from Portuguese customers. This classification might suggest some leakage since, when a hotel receives a booking and the origin of the customer is unknown, the hotel classifies it as coming from a Portuguese customer and only at check-in is the *Country* correctly filled. If the booking is cancelled, there is a high probability that the customer continues to be classified as Portuguese.

Figure 23 – H1 cancellation ratio per country



Further exploration of the datasets revealed additional insights into the hotels' operations. Analysis of the cancellation ratios per month (Figure 24) confirms the analysis of the cancellations ratios per year (Figure 17): an increasing trend in cancellations ratio. However, as illustrated in Figure 24, cancellation ratios diverge quite significantly per hotel and month, particularly for H1 and H4. This divergence can be explained by the fact that both hotels were closed during some periods between 2013 and 2015. Because the closure was not communicated in advance, cancellations reach values of 100% in those months. On the other hand, apart from these exceptions, it seems to be common to all the hotels that the period of the year with the highest cancellation ratio is in the high season, more precisely July and August. Surprisingly, is not in the months of high demand that lead time and cancellation time seems to be higher. As shown in Figure 25, for H1 and H4 there are peaks associated with the time of closure, but there are others associated with special events that occurred in the region (e.g., December 2014 and January 2015).

Figure 24 - Cancellation ratio per month

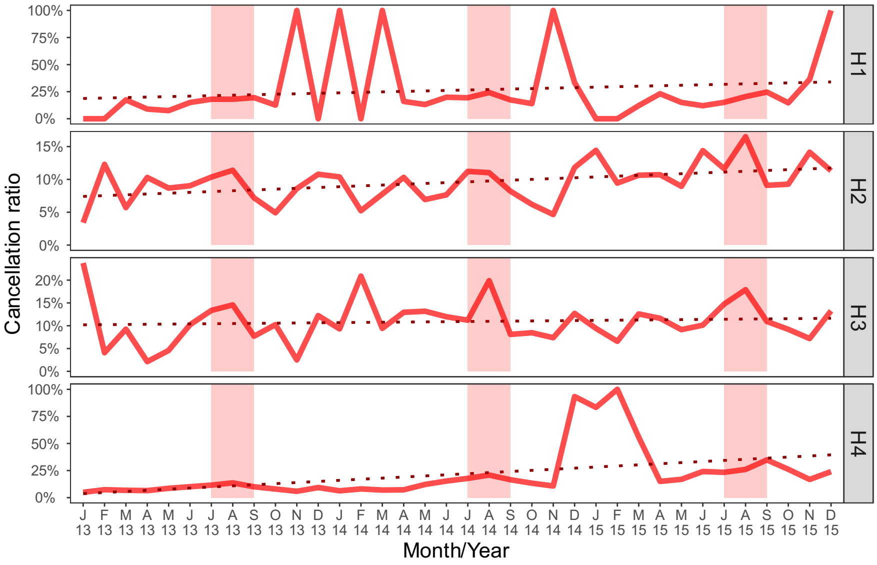
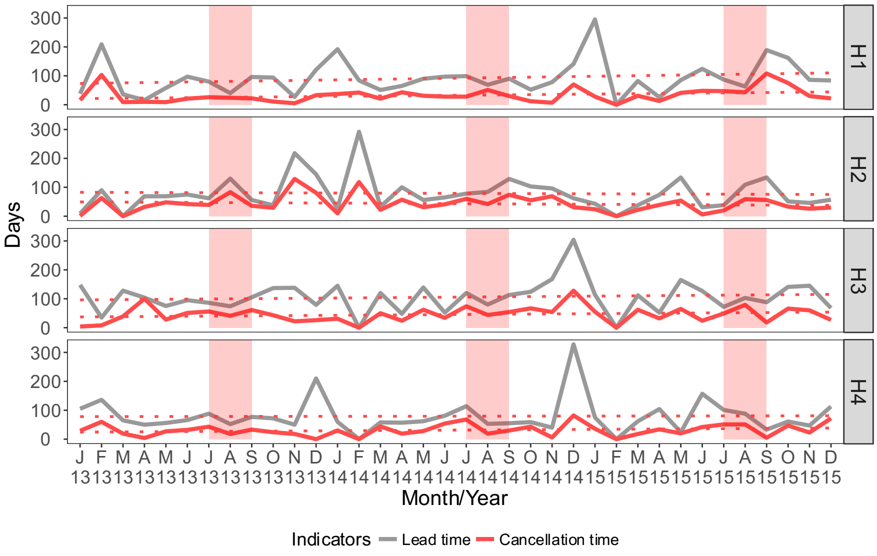


Figure 25 - Lead time and cancellation time per month



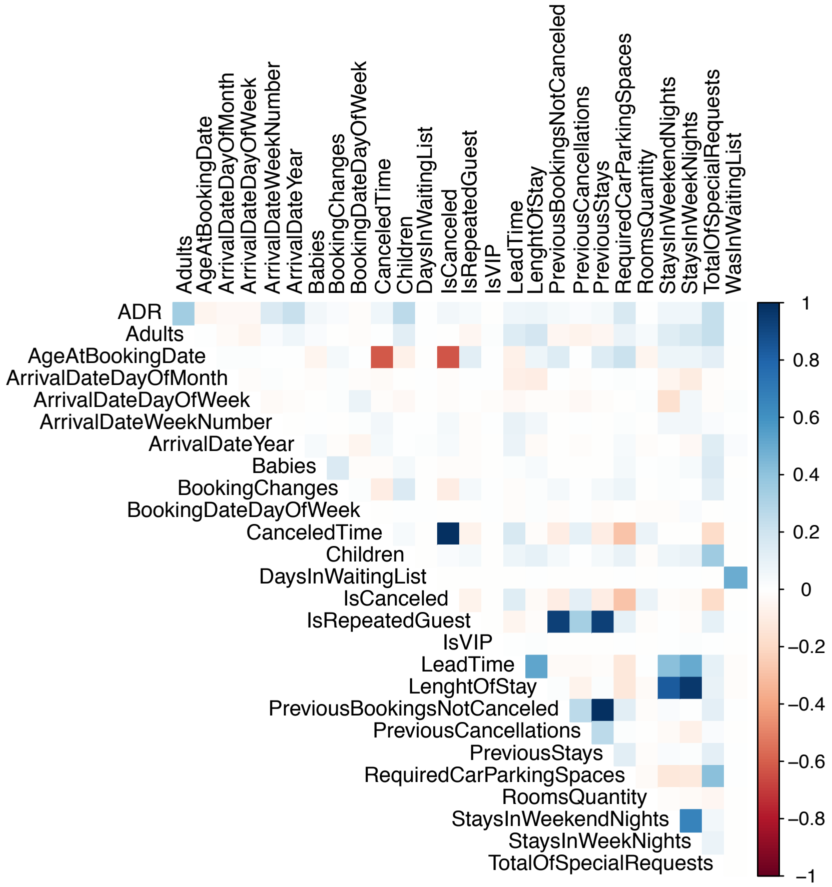
Summary statistics also showed that some features presented no values for some of the hotels. *RequiredCarParkingSpaces* was not used by H2 and H3, which is comprehensible since these two hotels do not have a garage or require of the customers to inform on how many car parking spaces do customers need. *MarketSegment*, *TotalOfSpecialRequests*, and *DaysInWaitingList* were other features that were empty or were filled with default values for some of the hotels.

3.2.3 Data preparation

The data preparation phase covered all activities related to the construction of the dataset to be used for modeling (modeling dataset). As for data understanding, data preparation required several iterations with the following phases before it was possible to create the modeling datasets.

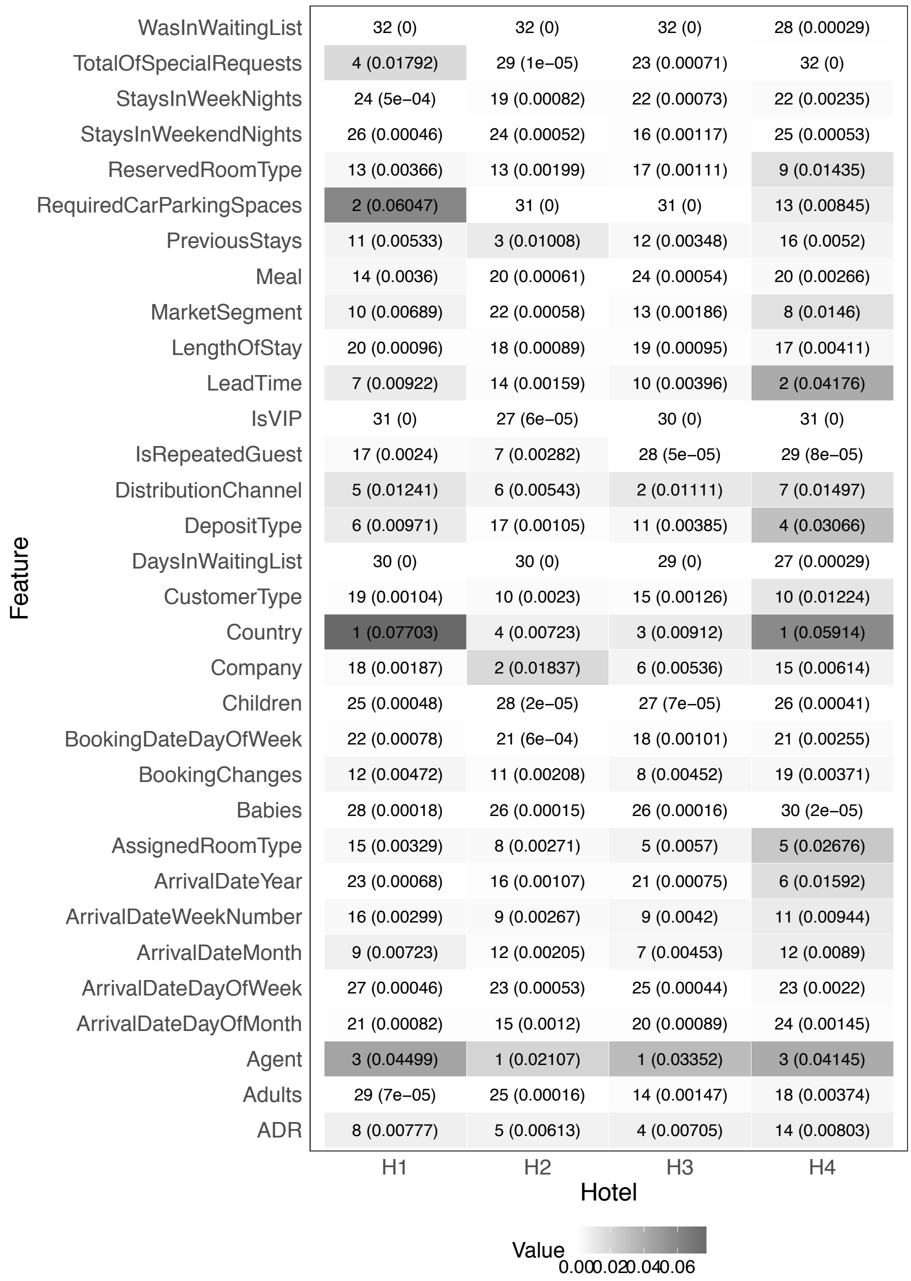
Feature selection is primarily focused on the removal of redundant or non-informative predictors (Guyon, Elisseeff 2003; Kuhn, Johnson 2013). In addition to the issues already identified during data exploration and data quality verification that pointed out features that could be removed from the modeling dataset, to select which ones could be removed features were ranked using two recommended methods for this task, correlation coefficient and mutual information (Guyon, Elisseeff 2003). First, the Spearman correlation coefficient between all numeric, integer and categorical features was studied (only categorical features that could be represented in the form of rank were included). Very high feature correlation does not signify the nonexistence of feature complementarity, but *"perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them"* (Guyon, Elisseeff 2003, p. 1164). Correlations among features were very similar in all datasets. As exemplified in the correlation plot of H1 dataset (Figure 26), some features presented high correlation coefficients with the outcome label (*IsCanceled*), which suggests leakage problems. This was the case of the features *AgeAtBookingDate* and *CanceledTime* that confirmed the findings concerning these features during the data understanding phase and showing that they should not be included in the modeling dataset. Naturally, *LengthOfStay* was highly correlated with *StaysInWeekendNights* and *StaysInWeekNights*. Since the latter features are more informative than *LengthOfStay*, in the sense that not only inform the model of the duration of the stay but also of the days of the week covered by the stay, *LengthOfStay* was also removed from the modeling dataset. Other features that were highly correlated between themselves were *IsRepeatedGuest*, *PreviousBookingsNotCanceled*, and *PreviousStays*. This high correlation is expected, as only repeated guests would have previous stays and previous bookings. However, previous bookings cancellations (feature *PreviousCancellations*), which is a feature only affected by repeated guests, was not highly correlated with these three. Thus, feature engineering was employed to create a feature, *PreviousCancellationRatio*, a ratio between previous cancellations (*PreviousCancellations*) and the sum of all previous bookings (*PreviousCancellations* + *PreviousBookingsNotCanceled*). This new feature, together with the results obtained from the evaluation of the models, showed that it was possible to remove the features *PreviousCancellations* and *PreviousBookingsNotCanceled*. Results from the modeling evaluation phase also showed that the feature *RoomsQuantity* could also be removed from the modeling datasets. More information about these features can be found in Appendix C.

Figure 26 – H1 Spearman correlation plot



Second, a mutual information filter was employed to confirm what was being shown by the results from the evaluation of models about other non-informative or noisy features that could be removed. Albeit tests have been made with other feature selection methods, including other filter selection methods (Pearson, Kendall, Chi-Squared and Spearman), the mutual information filter was chosen not only because of its proved adequacy (Guyon, Elisseeff 2003), but also, because filter methods are less expensive in computational terms and tend to overfit less then wrapper methods (Chandrashekar, Sahin 2014; Kuhn, Johnson 2013). Mutual information is a measure of dependence between two random variables (Cover, Thomas 1991). The mutual information filter assesses the contribution of a variable towards reducing uncertainty about a feature and the cancellation outcome label. Rank and value results of the mutual information filter for all hotels are shown in Figure 27. Results confirmed some of the assumptions made during data understanding and initial model evaluation phases, leading to the removal of additional features, such as *ArrivalDateDayOfWeek*, *BookingDateDayOfWeek*, and *PreviousStays*. The results also confirmed that features had different predictive relevance from one hotel to another, and that some features, although not relevant for some of the hotels, were relevant for others, like the features *TotalOfSpecialRequests*, *RequiredCarParkingSpaces*, *DayInWaitingList* or ArrivalDateYear. Mutual information filter results also confirmed that some features had almost no importance to reduce uncertainty in cancellations, namely *WasInWaitingList* and *IsVIP*. This fact resulted in the removal of these features from the modeling dataset. Nevertheless, other low raking features, such as *StaysInWeekNights* and *StaysInWeekendNights*, were not removed from the datasets because, as acknowledged by Guyon, Elisseeff (2003), sometimes features that are useless by themselves can provide significant performance improvement when used in conjunction with other features. Since these features were the only ones that represented the duration of the stay, it was decided to keep them in the datasets.

Figure 27 - Mutual information filter results (rank and value)



Feature	H1	H2	H3	H4
WasInWaitingList	32 (0)	32 (0)	32 (0)	28 (0.00029)
TotalOfSpecialRequests	4 (0.01792)	29 (1e−05)	23 (0.00071)	32 (0)
StaysInWeekNights	24 (5e−04)	19 (0.00082)	22 (0.00073)	22 (0.00235)
StaysInWeekendNights	26 (0.00046)	24 (0.00052)	16 (0.00117)	25 (0.00053)
ReservedRoomType	13 (0.00366)	13 (0.00199)	17 (0.00111)	9 (0.01435)
RequiredCarParkingSpaces	2 (0.06047)	31 (0)	31 (0)	13 (0.00845)
PreviousStays	11 (0.00533)	3 (0.01008)	12 (0.00348)	16 (0.0052)
Meal	14 (0.0036)	20 (0.00061)	24 (0.00054)	20 (0.00266)
MarketSegment	10 (0.00689)	22 (0.00058)	13 (0.00186)	8 (0.0146)
LengthOfStay	20 (0.00096)	18 (0.00089)	19 (0.00095)	17 (0.00411)
LeadTime	7 (0.00922)	14 (0.00159)	10 (0.00396)	2 (0.04176)
IsVIP	31 (0)	27 (6e−05)	30 (0)	31 (0)
IsRepeatedGuest	17 (0.0024)	7 (0.00282)	28 (5e−05)	29 (8e−05)
DistributionChannel	5 (0.01241)	6 (0.00543)	2 (0.01111)	7 (0.01497)
DepositType	6 (0.00971)	17 (0.00105)	11 (0.00385)	4 (0.03066)
DaysInWaitingList	30 (0)	30 (0)	29 (0)	27 (0.00029)
CustomerType	19 (0.00104)	10 (0.0023)	15 (0.00126)	10 (0.01224)
Country	1 (0.07703)	4 (0.00723)	3 (0.00912)	1 (0.05914)
Company	18 (0.00187)	2 (0.01837)	6 (0.00536)	15 (0.00614)
Children	25 (0.00048)	28 (2e−05)	27 (7e−05)	26 (0.00041)
BookingDateDayOfWeek	22 (0.00078)	21 (6e−04)	18 (0.00101)	21 (0.00255)
BookingChanges	12 (0.00472)	11 (0.00208)	8 (0.00452)	19 (0.00371)
Babies	28 (0.00018)	26 (0.00015)	26 (0.00016)	30 (2e−05)
AssignedRoomType	15 (0.00329)	8 (0.00271)	5 (0.0057)	5 (0.02676)
ArrivalDateYear	23 (0.00068)	16 (0.00107)	21 (0.00075)	6 (0.01592)
ArrivalDateWeekNumber	16 (0.00299)	9 (0.00267)	9 (0.0042)	11 (0.00944)
ArrivalDateMonth	9 (0.00723)	12 (0.00205)	7 (0.00453)	12 (0.0089)
ArrivalDateDayOfWeek	27 (0.00046)	23 (0.00053)	25 (0.00044)	23 (0.0022)
ArrivalDateDayOfMonth	21 (0.00082)	15 (0.0012)	20 (0.00089)	24 (0.00145)
Agent	3 (0.04499)	1 (0.02107)	1 (0.03352)	3 (0.04145)
Adults	29 (7e−05)	25 (0.00016)	14 (0.00147)	18 (0.00374)
ADR	8 (0.00777)	5 (0.00613)	4 (0.00705)	14 (0.00803)

Data cleaning and data transformation are some of the other tasks involved in data preparation (Abbott 2014; Kuhn, Johnson 2013). Some features presented small quality issues, like high positive or negative skewness (e.g., *LeadTime*) or outliers (e.g., *ADR*). Therefore, transformations functions (e.g., Log10 in *LeadTime*) were tested on them. However, the evaluation results showed that, in general, the models' performance did not improve by using these transformations.

3.2.4 Modeling

Due to the differences found in hotels' data, namely the differences in categorical features levels and the order and magnitude of features' contribution to the outcome, a model was built for each of the hotels. Since the outcome label (*IsCanceled*) only assumes binary values (0: not canceled; 1: canceled), to assess which algorithms performed better initial models were built using all two-class classification algorithms available in Microsoft Azure Machine Learning Studio. Given that initial results from the algorithms Average Perceptron, Bayes Point Machine and Logistic Regression were far worse than the results of other algorithms, subsequent models were only built using Boosted Decision Tree (BDT), Decision Forest (DF), Decision Jungle (DJ), Locally Deep Support Vector Machine (LDSVM) and Neural Network (NN).

Cross-validation was used to evaluate the performance of each one of the models, specifically *k*-fold cross-validation, a well-known and widely used model assessment technique (Hastie, Tibshirani, Friedman 2001). Although cross-validation can be computationally costly (Smola, Vishwanathan 2008), it allows for the development of models that are not overfitted and can be generalized to independent datasets. K-fold cross-validation works by randomly partitioning the sample data into *k* sized subsamples. In this case, data was divided in 10 folds – a typical number of chosen folds (Hastie, Tibshirani, Friedman 2001; Smola, Vishwanathan 2008). Then, each of the 10 folds was used as a test set and the data in the remaining 9 as training data. Performance measures were calculated for each of the ten folds, for which mean and standard deviation were calculated to assess the global performance of each algorithm. R scripting was used in the computation and presentation of these two measurements. A high-level diagram of the 10-fold cross validation process is shown in Figure 28. Table 3 presents the results for each of the five employed algorithms.

The classification result is a continuous value between 0 and 1. It is the cutoff or threshold that defines to which class the outcome should be assigned. A standard fixed threshold of 0.5 was used, meaning results below 0.5 were classified as 0 (non-canceled) and all others as 1 (canceled).

Cross-validation results were auspicious. In all hotels, the lowest *Accuracy* mean result was 0.879, registered for H1 using the neural network algorithm, while most models reached mean *Accuracy* values above 0.9. If *AUC* is taken as the assessment measure, all models, independently of the hotel, presented values above 0.9. Standard deviation values also shown that there was low variance among the models could be generalized to other datasets of the same hotel.

Regarding *Accuracy*, DF achieved the highest scores. Regarding *Precision*, DF was also the best for three out of four hotels. BDT presented slightly lower values regarding *Accuracy* and *Precision* but was the best model for three out of the four hotels regarding the other measures (*Recall*, *F1Score*, and *AUC*). Hence, optimized DF and BDT models were built to assess their performance.

Figure 28 – High-level visualization of the 10-fold cross-visualization procedure

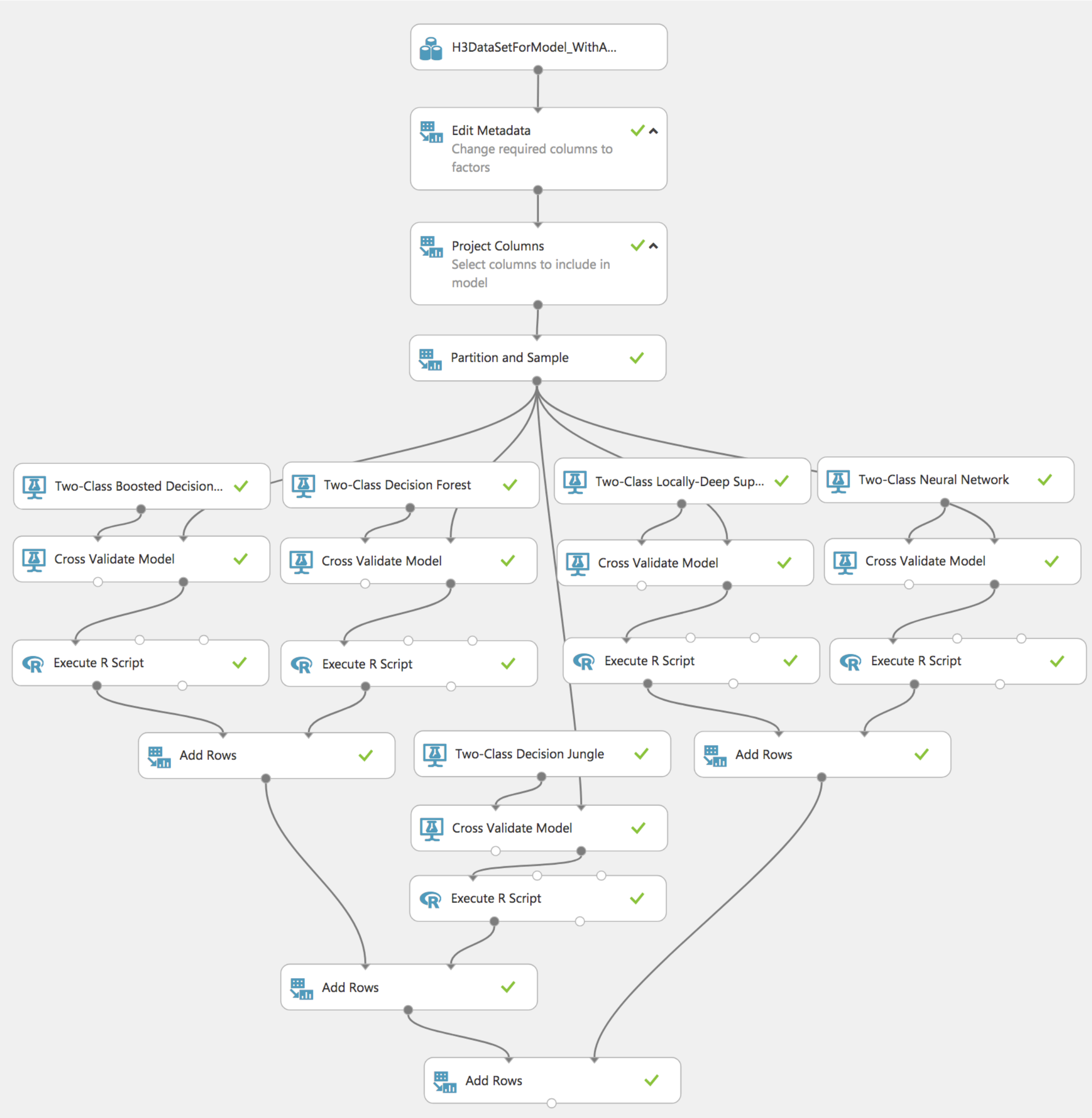


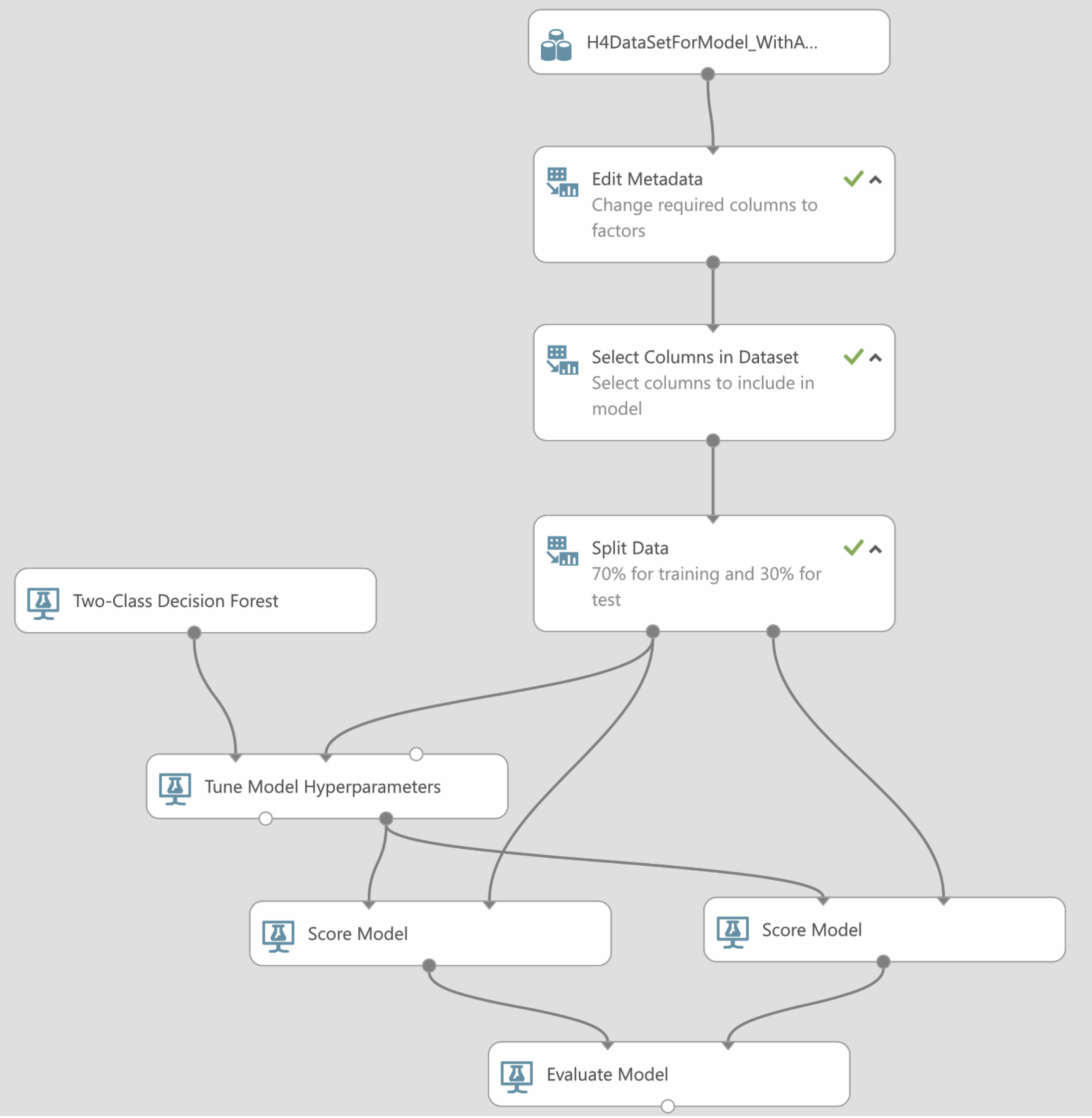
Table 3 - 10-fold cross-validation results

Hotel	Algorithm	Measure	Accuracy	Precision	Recall	F1 Score	AUC
H1	BDT	Mean	0.907	0.767	0.671	0.716	0.943
		SD	0.003	0.015	0.022	0.013	0.003
	DF	Mean	0.908	0.817	0.611	0.699	0.933
		SD	0.004	0.015	0.020	0.016	0.004
	DJ	Mean	0.882	0.953	0.340	0.501	0.906
		SD	0.004	0.025	0.021	0.024	0.009
	LDSVM	Mean	0.892	0.853	0.463	0.599	0.904

Hotel	Algorithm	Measure	Accuracy	Precision	Recall	F1 Score	AUC
		SD	0.006	0.039	0.031	0.029	0.008
	NN	Mean	0.879	0.664	0.637	0.646	0.911
		SD	0.007	0.058	0.063	0.014	0.006
H2	BDT	Mean	0.983	0.930	0.898	0.913	0.976
		SD	0.003	0.028	0.034	0.018	0.014
	DF	Mean	0.983	0.960	0.873	0.914	0.968
		SD	0.005	0.027	0.045	0.028	0.017
	DJ	Mean	0.982	0.955	0.860	0.904	0.980
		SD	0.003	0.027	0.039	0.018	0.011
	LDSVM	Mean	0.983	0.954	0.871	0.910	0.953
		SD	0.003	0.023	0.030	0.019	0.017
	NN	Mean	0.976	0.888	0.877	0.882	0.967
		SD	0.004	0.034	0.030	0.020	0.008
H3	BDT	Mean	0.972	0.894	0.861	0.877	0.965
		SD	0.004	0.026	0.027	0.018	0.011
	DF	Mean	0.973	0.938	0.822	0.876	0.947
		SD	0.003	0.015	0.029	0.019	0.014
	DJ	Mean	0.972	0.911	0.843	0.876	0.962
		SD	0.003	0.024	0.017	0.015	0.009
	LDSVM	Mean	0.970	0.930	0.806	0.864	0.934
		SD	0.003	0.019	0.020	0.018	0.011
	NN	Mean	0.960	0.838	0.822	0.829	0.942
		SD	0.007	0.056	0.029	0.027	0.013
H4	BDT	Mean	0.927	0.802	0.705	0.750	0.952
		SD	0.005	0.013	0.035	0.024	0.006
	DF	Mean	0.928	0.835	0.672	0.744	0.948
		SD	0.004	0.020	0.027	0.019	0.006
	DJ	Mean	0.898	0.833	0.443	0.567	0.924
		SD	0.010	0.057	0.105	0.094	0.008
	LDSVM	Mean	0.915	0.814	0.590	0.684	0.919
		SD	0.006	0.033	0.024	0.023	0.004
	NN	Mean	0.907	0.710	0.680	0.694	0.932
		SD	0.006	0.029	0.035	0.020	0.007

As usual when creating machine learning predictive models, as depicted in the high-level visualization of the modelling procedure for DF algorithm (Figure 29), the datasets were divided in two stratified subsets, one using 70% of data for training (model learning) and another with the remaining 30% to test the developed model.

Figure 29 - High level visualization of DF modeling procedure


H4DataSetForModel_WithA...
Edit Metadata
Change required columns to factors
Select Columns in Dataset
Select columns to include in model
Split Data
70% for training and 30% for test
Two-Class Decision Forest
Tune Model Hyperparameters
Score Model
Score Model
Evaluate Model

Model parameters were optimized by applying the function “Tune model hyperparameters” to the training set thus testing different combinations of each algorithm’s parameters, and with that, determine the optimum parameters to use. Parameters tuning was made in five random sweep runs, using the metric *F1Score* to assess the performance of the parameters.

3.2.5 Evaluation

Test results of the models built with BDT and DF algorithms are presented in Table 4. Regarding *Accuracy*, BDT presented higher or equal values to DF. In terms of *F1Score*, BDT also presented the highest results in three out of the four hotels. By contrast, in terms of *AUC*, DF presented higher results in three out of the four hotels. Although both models present slight differences, overall performance is comparable. For H2 and H3, both reach *Accuracy* values above 0.97 and *AUC* values above 0.96. For H1 and H4, results were lower but, nonetheless, outstanding values.

Still, another important metric to consider is the number of false positives, in particular, if the hotel decides to use prediction results to contact bookings identified as likely to cancel. If this is the case, the smaller number of false positives the model generates, the least the hotel will spend in compensations with bookings that would turn out not to be canceled. If this is considered, the DF algorithm should be chosen as the one to use, as its model presents the lower number of false positives for the hotels' sets. These results seem to validate the findings of Fernández-Delgado, Cernadas, Barro, Amorim (2014). These authors tested 179 classifiers from 17 families and concluded that the best results are usually obtained with the random forest algorithms family.

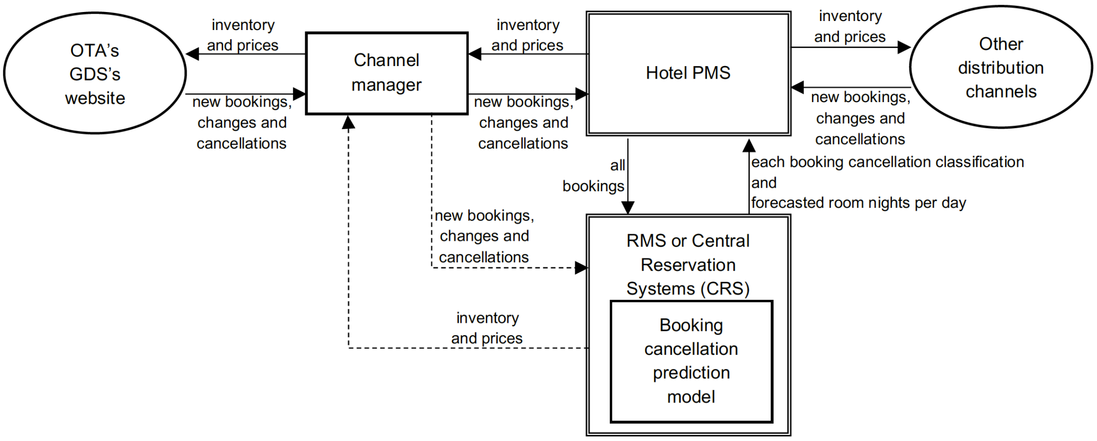
Table 4 - Optimized BDT and DF models

Hotel	Algorithm	TP	FP	FN	TN	Accuracy	Precision	Recall	F1 Score	AUC
H1	BDT	679	131	379	4 907	0.916	0.838	0.642	0.727	0.936
	DF	541	94	517	4 944	0.900	0.852	0.511	0.639	0.935
H2	BDT	259	11	31	2 629	0.986	0.959	0.893	0.925	0.974
	DF	255	5	35	2 635	0.986	0.981	0.879	0.927	0.977
H3	BDT	285	35	38	2 451	0.974	0.891	0.882	0.886	0.963
	DF	272	22	51	2 464	0.974	0.925	0.842	0.882	0.971
H4	BDT	1 120	270	430	8 153	0.930	0.806	0.723	0.762	0.940
	DF	1 000	220	550	8 203	0.923	0.820	0.645	0.722	0.948

3.2.6 Deployment

Despite the fact that the deployment of these models in a production environment is not in the scope of this chapter, models' deployment is critical to assess their success. Based on was learned from the construction of the models, it is now possible to define a framework for the deployment of the models. The booking cancellation prediction model should not be implemented by itself (Figure 30). In truth, if deployed independently of the hotel's remaining systems, it is unlikely that it would present any valid results in terms of revenue management. Today's speed and complexity imposed on a hotel reservations department are such that advantages of using the model could not be apparent if tasks related to the model inputs and outputs had to be performed by hand. For example, defining prices and inventory to publish in online platforms based on demand forecasts is something that is very difficult to be done without automatic help, at least, in a timely fashion. Thus, the model should be integrated in the hotel RMS or, eventually, in the hotel chain CRS (Central Reservation System). This integration would enable the system to achieve more accurate net demand forecasts and consequently, present better overall forecasts.

Figure 30 - Model deployment framework


OTA's
GDS's
website
inventory and prices
new bookings, changes and cancellations
Channel manager
inventory and prices
new bookings, changes and cancellations
Hotel PMS
inventory and prices
new bookings, changes and cancellations
Other distribution channels
all bookings
each booking cancellation classification and forecasted room nights per day
new bookings, changes and cancellations
RMS or Central Reservation Systems (CRS)
inventory and prices
Booking cancellation prediction model

By being directly connected to the PMS, the RMS/CRS can pass to the PMS the adjusted inventory. This inventory could then be communicated by the PMS automatically (or by CRS as sometimes happens), directly or via a channel manager, to the different distribution channels (OTA's, GDS's, travel operators, hotel website, among others). This automation of the inventory allocation based on better net demand forecast enables the hotel to instantly react, in case of a booking cancellation or in case of a change in a booking cancellation classification, adjust its sale inventory and communicate it to the different distribution channels.

Models elaboration drew the attention to other factors that needed to be considered when deploying the models. First, some predictor features change with time (e.g., *LeadTime*) or can assume new values every day, as in the case of changes/amendments to bookings (e.g., *BookingChanges* or *Adults*). Thus, the model should be run every day so that all in-house bookings and results can be evaluated on a daily basis. Second, as stressed by Abbott (2014, p. 618) *"even the most accurate and effective models don't stay effective indefinitely. Changes in behavior due to new trends, fads, incentives, or disincentives should be expected"*. For example, temporary hotel closing at different times in the year, as it happened with H1 or H4, or when a hotel changes its marketing efforts and starts to capture more market from OTAs in substitution of traditional tour operators, will influence many predictor features, such as *MarketSegment*, *DistribuitonChannel* and *LeadTime*. If the model is not updated, its performance will decline.

3.3 Discussion

The results achieved far exceeded the initially established objective of 0.8 of *Accuracy* and 0.8 of *AUC*. The results unquestionably demonstrate that, despite what was alleged by Morales, Wang (2010), features extracted and derived from the hotels' PMS databases are a good source to predict with high accuracy if bookings are going to be canceled. *Accuracy* reached 0.99 in H2 and values above 0.9 for the remaining hotels. *AUC* was consistently superior to 0.93, which is considered "excellent" (Zhu, Zeng, Wang 2010). In general, the results are superior in terms of *Accuracy* to the results recently obtained for the same type of problem by van Leeuwen (2018), while in line with its *Precision* results. However, the same did not happen with *Recall* and *F1Score*. Compared to the results obtained by Huang, Chang, Ho (2013) for the same type of problem but for the restaurant industry, these results are clearly superior in terms of *AUC*, *Accuracy,* and *Precision*. On the other side, very good performance results may suggest overfitting[6] or leakage problems, which raises the question if this level of results could be maintained in a production environment (Abbott 2014).

An important part of the time spent building forecast/prediction models consists in collecting and preparing the necessary data (Talluri, Van Ryzin 2005). Suitable data and a good selection of features are crucial for models' performance. As mentioned earlier and illustrated by Figure 27, not all features have the same order of importance, nor do they contribute the same to predict if a booking is going to be canceled. This calls for a specific characterization from each hotel. Hotel location, services, facilities, the nationality of guests, markets, and distribution channels are among the many features with different weights for predicting cancellation. One example is the feature *RequiredCarParkingSpaces*. It is ranked in second place for H1 and 13th for H4 but with no importance in terms of H3 and H4. This low importance is easily understandable if one knows these hotels' operations that do not have such limited car parking spaces as H1 and H4. Therefore, hotel revenue management and general business domain knowledge are not enough to undertake a good selection of features. It is also essential to understand each hotel's operation modes and characteristics. This understanding can make a difference in terms of final model performance and adequacy. For this reason, hotel prediction modeling should use detailed booking data from the hotels' PMS, in counterpoint to data in more strict formats like the PNR format.

As with any other predictive analytics problem, developing a model to predict booking cancellations requires data that meet all of the attributes of quality data: accurate, reliable, unbiased, valid, appropriate, and timely (McGuire 2017; Rabianski 2003). As previously mentioned, some of the datasets features had outliers (e.g., *ADR* for H1 dataset). Lack of quality can affect model performance, and hotels that want to build prediction models should ensure that they have an adequate data quality policy in place.

Besides producing demand forecasts that can be aggregated at different levels (globally, by distribution channel, by market segment, among others), classification prediction models have an advantage that regression prediction models do not possess: they allow hotel managers to take action on bookings identified as likely to cancel. To avoid potential cancellations, hotel managers

[6] In statistics and machine learning, the term "overfitting" is used to describe a situation where the model corresponds too well to the training data, but fails to generalize to unseen data, thus not predicting reliably the result of future observations (Hastie, Tibshirani, Friedman 2001).

can contact customers prior to their expected arrival date and offer services, discounts, entrances to shows/amusement parks, or other perks. Understandably, these measures cannot be applied to all customers since some are known to be insensitive to these kinds of offers (e.g., corporate guests). Moreover, due to the direct influence of forecast accuracy in the performance of revenue management, the implementation of these booking cancellation prediction models, in the context of a revenue management system framework, as depicted in Figure 30, could represent a significant contribution to reduce uncertainty in the inventory allocation and pricing decision process.

For studied hotels, cancellations from 2013 to 2015 exceeded an amount of 6.2 million euros. Of course, not all this amount is lost revenue, as new bookings replace many cancellations. Nonetheless, if models' predictions can be used to prevent some of these cancellations, even if only a small fraction of them like 10%, models could have a significant impact in terms of revenue performance. Not only would revenue directly increase due to the avoidance of cancellations, but also because it would allow revenue managers to be more assertive in pricing and inventory allocation decisions.

3.4 Summary

CRISP-DM process model revealed to be an adequate method to build exploratory bookings cancellation prediction models. Although requiring multiple iterations between the different phases, the main objectives previously outlined for the chapter were achieved: understanding the problems related to data collection, data quality, data preparation, and modeling. This understanding and the understanding of the limitations is fundamental for developing final models and to study how models could be deployed in a production environment.

The use of highly detailed booking data from four resort hotels' PMS, from the period of 2013 to 2015, with cancellation rates spanning from 8.8% to 26.4%, proved to be a good choice to understand the similarities and dissimilarities between hotels and to build the models.

Data visualization and data mining techniques, together with summary statistics and the mutual information filter demonstrated to be good tools to understand data patterns, trends, and anomalies. The resulting analyses showed the differences in hotel operations, how some hotel data have more outliers than others, and how some features seem to have similar predictive power in all hotels while others do not.

The combination of local and cloud resources, namely the use of R and Microsoft Azure Machine Learning, allowed for the evaluation of multiple machine learning algorithm models and the conclusion that decision forest algorithms were the ones that better suited this type of problem and data.

The good performance results are a good indication that RQ1: "Could a booking's cancellation prediction model that uses PMS data display better results than a model that uses PNR data?", can be answered affirmatively. At the same time, results also raised new questions, such as if similar performance could be achieved with other hotels' data, in particular with another type of hotels (city hotels instead of resort hotels), what kind of improvement could data from additional sources bring to the performance of models, or if this level of results could be achieved in a production environment. Answering these questions and confirming the positive answering to RQ1, will be the subject of the following chapters.

4 FINAL MODELS

The results obtained by the exploratory models confirmed that combining data science tools and capabilities, namely data mining, machine learning, and data visualization, with highly detailed data extracted from hotels' PMS, makes it possible to build models capable of predicting cancellation of bookings. For confirming that highly detailed PMS data can indeed produce good prediction results, new models were built using data from eight hotels: four resort and four city hotels. The new models address the limitations found in the exploratory models, such as overfitting and features' leakage of the outcome. Additionally, data from other sources, such as weather, social reputation, and competitors' prices are to be used to study what is its impact on the improvement of the models' performance. Furthermore, prediction models will be used in the true meaning of the word "prediction" as in understanding the past, that is, explain which are the features that influence the probability of canceling, or in other words, which are the drivers for the cancellation of a booking. The new models will enable an answer for RQ1 - if models built with PMS data could produce better results than models built with data in the PNR format - and RQ2 - if models could be improved with the inclusion of data from additional sources. At the same time, the development of the new models provide clues as for how can these models be deployed in a real production environment or RQ3.

After an introductory section where the impact of big data in forecasting is addressed, a detailed description of the methods and materials employed for the elaboration of the models is presented. The influence in the research will also be discussed and the chapter ends with a summary and a preview of the work needed to answer the third research question.

4.1 Introduction

CRISP-DM, due to its appropriateness to the problem, as confirmed in the previous chapter, was again selected as the process model to be applied in the development of the final models which were built for eight different hotels. The rationale behind the decision to increment the number of hotels to be studied is to learn if the patterns for cancellations were found in different types of hotels if a models' structure could be maintained for every hotel, and to perceive what are the similarities and dissimilarities between the hotels' operations. Thus, and because hotels' categories differ, data from two distinct types of hotel was employed: city and resort hotels. Among other differences, city hotels differ from resort hotels on demand by season and by the length of stay, as shall be confirmed. While the demand for city hotels is mostly stable during the year, for a resort hotel is highly seasonal. Moreover, a city hotel guest usually book short stays, while in resort hotels, at least in some periods, it is common to have stays for more than seven nights.

In terms of business, the hotel industry is not different from other industries, and business operations change over time. Therefore, booking's cancellation patterns and tendencies tend to change over time - this over-time non-stationary distribution of input features when in regard of the outcome label is known as "concept drift" (Gama, Medas, Castillo, Rodrigues 2004; Webb, Hyde, Cao, Nguyen, Petitjean 2016). Taking this matter into consideration, it was decided that data must be from the same period for every hotel. Based on this prerequisite, independent and small chain hotels which could provide data for a common period were contacted. From those, a total of eight, four city hotels and four resort hotels, agreed to provide access to copies of their PMS' databases for this research. Because hotels required anonymity (for them and their customers), henceforth city hotels will be identified from C1 to C4, and resort hotels from R1 to R4. Each of the hotels was asked to identify its five-hotel competitive set, that is, is *"a group of similar and directly competing lodging properties to which an individual hotel's operating performance is compared"* (Hayes, Miller 2011, p. 22).

One of the main differences between the final models and the exploratory models is the combination of PMS data with data from additional sources in some of the final models. As presented in Chapter 2, although advocated by several authors as beneficial for forecast performance, until now no works have combined PMS data with data from other sources in order to develop cancellation forecast/prediction models. Variety, i.e., the use of multiple data sources and different data types (structured and unstructured) is one of the characteristics of "big data". The other two characteristics are volume and velocity (Günther, Rezazade Mehrizi, Huysman, Feldberg 2017; McGuire 2017; Wang 2015). Although research on the application of big data in tourism and hospitality fields is still scarce (Pan, Yang 2017a), several interesting examples already exist that demonstrate its potential. For example, Pan and Yang (2017b) used search engine queries, website traffic, and weather data to forecast hotel occupancy. Song and Liu (2017) presented a framework for predicting tourism demand. Liu, Teichert, Rossi, Li and Hu (2017) employed big data to investigate language-specific drivers of hotel satisfaction. Kahn and Liu (2016) showed how electricity big data could be used to help hotels improve energy efficiency.

While PMS data should capture some of the factors that influence demand and cancellations (like time to arrival, customer segment, duration of stay, season of stay or cancellation policy), data from non-PMS sources should capture other possible influential factors, such as social reputation, currency exchange rates, weather, and competition (McGuire 2016; Talluri, Van Ryzin 2005). Defining which sources could provide data bearing this type of information and then extract, store, and process the data proved to be one the most demanding tasks. The selection of data sources was decided by the criteria that data sources should hold data that could represent any of the previously mentioned factors or any other that could explain why customers cancel their bookings or do not show up. Based on these criteria, PMS data was merged with data from national and local holiday calendars, local weather forecasts, special events calendars, currency exchange rates, stock exchange indexes, social media reputation (including those of the hotels' direct competitors), and online prices/inventory availability for future dates (also including those of the competitor sets). In fact, the identification of data sources to obtain this data can be a challenging assignment, as recognized by McGuire (2017). Data sources should meet two essential requirements: 1) disclose quality data, and 2) should when applied, capture the bi-dimensionality of hotel demand forecast. This bi-dimensionality is due to the need of having data to represent both the date for the creation of the booking and the date for the outcome (either arrival or cancellation date) (Weatherford, Kimes 2003). Taking weather as an example, despite its importance to explain hotel demand (McGuire, 2017 and Pan & Yang, 2017a), the incorporation of a weather forecast for far-off future dates is nonviable. However, depending on the data point selected, weather forecasts can be used as a feature in a machine learning model. This data point is the arrival date for bookings that are not canceled, or the cancellation date for canceled bookings. In this way, the model can use this feature to understand if the weather forecast is related to the booking cancellation outcome. Lastly, selected data sources had to be public and available for general use so that our work could be replicated and eventually applied by hotels. This meant that access to data had to be free and that extraction could be accomplished using the data providers' Application Programming Interfaces (APIs), or via web scraping. Based on these requirements, weather data was extracted from the Weather Underground website (Weather Underground [no date]). This popular website has a powerful API that allows one to obtain current and 10-day forecast weather conditions for almost anywhere in the world. To understand if stays covering a holiday show a different pattern of cancellations from stays not including holidays (for instance, find out if customers who take advantage of "long weekends" tend to cancel more when it rains than other customers), national and local holidays were extracted from the TimeAndDate.com website. TimeAndDate.com is considered to be the biggest time zone-related website (TimeAndDate.com [no date]). For data in the special events due to occur in the hotels' region, the selected source was the website (Lanyrd.com [no date]). The objective to gather special events data was to build features that could capture cancellation patterns that could be linked to events. Social reputation is today one of the main aspects influencing a customers' booking decision process (Anderson 2012; Cantallops, Salvi 2014; McGuire 2016; Viglia, Minazzi, Buhalis 2016). To understand if a change in a hotel social reputation could influence cancellations, online reviews from two of the most popular websites in the area were selected as sources for this data: Booking.com and Tripadvisor.com (European Commission 2014). To perceive if a change in price or in rooms' availability of a competitor hotel could make a “deal-seeking” customer to cancel a booking, prices and rooms available from the studied hotel and the hotel’s competitors were necessary. Booking.com was chosen as the source for this data due to its predominance in Europe, which contributes heavily for the influence Online Travel Agencies (OTA’s) exert on hotels (HOTREC - Association of Hotels, Restaurants and Cafes and similar establishments of Europe 2016; Martin-Fuentes, Mellinas 2018). Currency exchange rates data and stock exchanges indexes data was also extracted. The former from Apilayer (Apilayer [no date]), and the latter from the Wall Street Journal website (Wall Street Journal [no date]). The rationale behind the selection of these two sources was the need to build features capable of capturing macroeconomic effects in the cancellations, such as deterioration of economic conditions in the country of origin of the customers. However, due to reasons later explained, data from these data sources, although collected, were not employed.

Considering the timeframe required to extract data from the above-mentioned non-PMS data sources, it was decided to define the period for the study beginning at January 1st, 2016 and ending in November 30th, 2017 (later shortened to November 20th, 2017). Nonetheless, not all of the models would hold data from this entire period. Complications with the data collected from some of data sources, together with the need to evaluate the performance of models that used only PMS data versus models that used data from multiple sources, lead to the development of four different models. The first model used exclusively PMS features with arrivals from January 1st, 2016 to November 20th, 2017 - Model 1. A second model, again using PMS based features solely but with arrivals from a shorter period, 1st of August, 2016 to November 20th, 2017, - Model 2. The objective for using models with the same features and structure but with fewer observations was to understand if the reduction in the number of observations had a severe impact on the model’s performance. The third model - Model 3 - included features from all the sources (PMS, weather forecast, social media reputation, holidays, special events, and online prices/inventory), with observations from the same period as Model 2. The objective now was that of realizing if the inclusion of features from additional sources improved the performance. Lastly, an optimized model was specifically built for hotels R1 and C1 to verify if the inclusion of additional features related to how hotels operations and services reflect in the models - Model 4. The choice of these two was because they were the only that shared characteristics allowing the creation of the additional features. The period of observations for Model 4 was the same as for Models 2 and 3.

How data was collected, including how data extractors were built for obtaining data from the non-PMS data sources, its challenges and its difficulties are explained in detail in the following section, throughout the description of the different CRISP-DM process model phases.

4.2 Process model

Although a comprehensive description of each CRISP-DM process model phases is presented sequentially in the following sections, it implied an iterative process, following the earlier development of the exploratory models.

4.2.1 Business understanding

Hotels were classified from three-star to five-star, ranging in size from 86 to 230 rooms. R2 and R4 are mixed-ownership units. Contrarily to what happened with the exploratory models, none of the hotels was closed during the period of study. As illustrated in Figure 31 and detailed in Table 5, from January 1st, 2016 to November 20th, 2017, booking's cancellation ratios vary between 12.2% for R3, to 40.0% for C1. Except for the case of R1, which presents a cancellation ratio slightly superior to that of C4, cancellations ratios for city hotels are clearly higher to the cancellation ratios for resort hotels. The reason seems to be linked with the distribution channels. Unlike city hotels, for resort hotels traditional tour operators still, represent an important component of the distribution. The great exposition of city hotels to bookings issued by OTAs tends to favor "deal-seeking" customers. OTA's booking share for the eight studied hotels ranged from 4.5% to 83.2% (Figure 31), revealing a moderate correlation between that share and the booking's cancellation ratio (0.5255).

Figure 31 - OTA's bookings share vs cancellations ratio

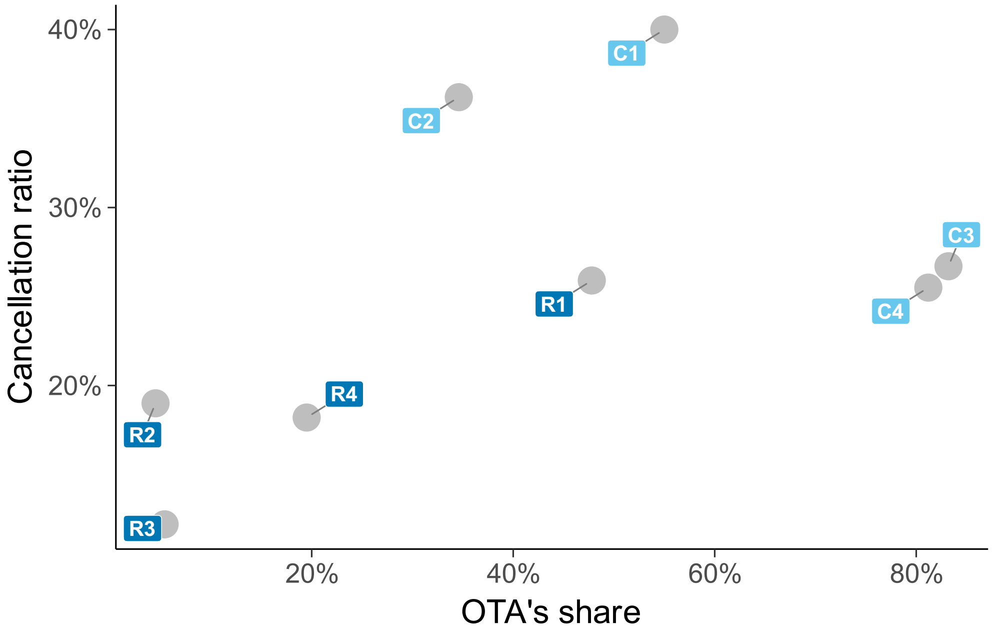


Table 5 - Hotels' cancellations and OTA's bookings share summary

Measure	C1	C2	C3	C4	R1	R2	R3	R4
Bookings not canceled	31 575	15 648	7 576	13 526	17 572	4 757	4 781	5 285
Bookings canceled	21 049	8 883	2 758	4 639	6 144	1 114	662	1 176
Cancellation ratio	40.0%	36.2%	26.7%	25.5%	25.9%	19.0%	12.2%	18.2%
OTA's share	55.0%	34.6%	83.2%	81.2%	47.8%	4.5%	5.4%	19.5%

Once again, the creation of the final models used a mix of local and cloud resources. However, due to the introduction of data in different formats and higher volume, the computational power and storage space needed for the final models intensively cloud-centric. The clear majority of the work happened using the Microsoft R server combined with Apache Hadoop and running on a HDInsight platform. The open version of R in this platform allowed us to take advantage of multi-processing and distributing computing to accelerate work tasks.

4.2.2 Data understanding

Given that multiple data sources were employed and that some of these required the development of data extractors to collect the data, in this chapter the sequence of presentation of the data understanding phase will be different from the previous chapter and referred to by data source.

4.2.2.1 Initial data collection

As previously introduced, so that the results per hotel could be comparable, the availability of hotels' PMS data outlined the selected period. Although some of the hotels' PMS data was accessible from the year 2012 onward, for others, data was only available from 2014 or 2015. Furthermore, because for most of the additional data sources historical data could not be extracted, it was decided to limit the period for the study ranging from January 1st, 2016 to November 30th, 2017. Because, to create features based in other data sources, the date at which the bookings outcome was known (canceled or not canceled) was needed, the bookings with a cancellation outcome date outside this period were excluded. During the defined period, non-PMS data extractors would have to run every day to collect data. Later on, due to some constraints in the extraction of non-PMS data, the study period has changed. Details of the data collection process and the reasons behind the change of the study period can be found in the following sections.

A list of features and summary statistics for each source dataset is available Appendix D. The description of each feature can be found in Appendix C.

4.2.2.1.1 PMS data

Again, as it happened with the exploratory models, despite the similarities between hotels' databases, specificities of each database and each hotel's particularities of operation were considered for building the data extraction SQL queries. However, this time, the process was facilitated by the findings of the exploratory models. In relation to exploratory models' datasets, nine features were removed (*AgeAtBookingDate*, *ArrivalDateDayOfWeek*, *BookingDateDayOfWeek*, *CanceledTime*, *IsVIP*, *LengthOfStay*, *PreviousStays*, *RoomQuantity*, and *WasInWaitingList*), and four new features were added (*ArrivalDateMonthYear*, *FolioNumber*, *ReservationStatus,* and *ReservationStatusDate*).

For hotels C1 and R1, two additional datasets were extracted. These regard a shorter period including expected arrivals from August 1st, 2016 instead of January 1st, 2016, but included eight other additional features (*AssocitatedToEvent*, *BookedSPA*, *SRDoubleBed, SRHighFloor*, *SRQuietRoom*, *SRTogether, SRTwinBed*, and *RateCode*). The feature *ArrivalDateMonthYear* was removed from these datasets. The datasets were collected to find out if the additional features could contribute for improving the model's performance (Model 4) and the reasons why will be explained later, with the description of the data preparation phase description.

The results of the initial evaluations with the training data were auspicious but, as usual, for unseen data the models did not perform as well, showing a tendency for overfitting. After some more iterations, the way data was being collected was changed to address this issue. Predictive modeling uses historical data to predict (so-called) future actions. For that to be effective, the timeline in historical data must be shifted. In other words, the values of input features should be acquired from a period prior to the fixation of the target variable (Abbott 2014). Booking data usually suffer changes and amendments, from the moment they are entered in the hotel PMS until the time of the guest check-out or cancels. Some of these changes and amendments intend to correct the information entered or to change the service required, including changing the period of stay, the number of persons, the type of meal, adding special requests or additional services (e.g., a SPA treatment). In fact, it is very common for hotels not to know certain details of the guest until check-in, including the country of origin, birthdate and other personal information. It is also common for guests to change their booking details at check-in time (e.g., add or remove more nights or change the number of persons). Understandably, some features' distributions differ per cancellation outcome. If the objective of the models is to predict bookings cancellation outcome, which is set at cancellation date or check-in date, the values of the input feature need to reflect these changes. So, instead of reading bookings' details directly from the PMS database reservations table, details were read from the PMS database "reservations change log" table. In other words, rather than read the last known booking details, the SQL queries had to be modified to read the details immediately before cancellation date or check-in date (according to the booking cancellation outcome).

4.2.2.1.2 Social media reputation data

Social media reputation is driven by user-generated content, including photos, videos, and reviews. Among all the sources of social media reputation, online reviews have long been recognized as one of the more credible sources of information. Customers often see themselves in the others' opinions, considering them as trustworthy (Leung, Law, Hoof, Buhalis 2013).

The importance of social media reputation data in our context deals with the need to capture the relation between the hotel reputation's rating and its competitive set's rating to understand if, when the former is lower than its competitors does influence the "deal-seeking" customers into canceling their bookings.

Booking.com and Tripadvisor.com, two of the most popular online reviews websites (European Commission 2014) were selected as the sources for collecting social media reputation data, in the form of online reviews. As many other websites who offer API's to access their data, both websites do not facilitate access to their APIs to academic researchers (Batrinca, Treleaven 2015). For this reason, a customized extractor was built for each of the websites. Extractors were built in C#, using Microsoft .Net Framework 4.5 and Selenium. Selenium browser automation tool enables navigation automation and content reading, thus allowing what is designated as web scraping or web harvesting (Batrinca, Treleaven 2015; Braun, Kuljanin, DeShon 2018). Extractors used Selenium together with Firefox browser to, in an automated way and in a daily basis, open and process the content of the web pages with reviews for each of the hotels and their competitors in the Booking.com and Tripadvisor.com websites. Besides collecting each hotel global information, such as the overall rating or the total number of reviews, detailed information about each new review was also collected (e.g., username, textual information, publication date, among others). Online reviews are originally unstructured data but, to be used in RMS, data needs to be structured. Having in consideration that, for online reviews data usage to be faster and more efficient, data should be stored in a relational database instead of a Hadoop environment (McGuire 2016), the extractors stored the processed data and the extraction metadata on a SQL Server database. Diagrams and dictionaries with detailed information of the database structure, including all metadata and data fields stored, as well as simple statistics of the collected data are available in Appendix E.

Although European law recognizes the right for users to make a copy of publicly available databases and their use in research (Bosch 2017; Monkman, Kaiser, Hyder 2018), it is common for companies to take measures that difficult this copy (scraping) (Jennings, Yates 2009). Example of these measures is the use of cookies, dynamic content generation via javascript or ajax, implementation of CAPTCHAs, rate limit requests, data obfuscation, malicious sources detection and blocking, among others (Imperva 2014), turning the development of web scraping extractors an increasingly difficult task. The first challenge is to ensure the necessary computational power is available, both in terms of storage and processing capacity (Batrinca, Treleaven 2015; Braun, Kuljanin, DeShon 2018). In our case, a Windows 2012 Server with a Xeon E3-1230 v3@ 3.30 Ghz CPU, 32 Gb of RAM, and 2 Tb of hard disk were used. Even with though, it took both extractors, around 5 to 8 hours daily to extract the global ratings and new reviews needed. The second challenge is that of constantly monitoring of the extractors to quickly react to changes in the website structure/content or the application of anti-scraping measures. For the present case, it consisted essentially in the form of ajax dynamic content which prompted users to select options being displayed, web pages' structure that differed in terms of the selected language, random displaying of pop-up overlays which required a click to allow page navigation, permanent changes on the website structure/content, and A/B testing. Responding to A/B testing, in particular, can be very demanding since the objective is to conduct a randomized experiment of showing two web pages variants to test which performs better (Kohavi, Longbotham 2017), it requires extractors to be adapted regularly and recognize both versions of test web pages. In the case of Booking.com, during the data collection period, more than 20 A/B tests were carried just on the online review's web pages.

For reducing storing and processing requirements and for facilitating data understanding and data preparation, SQL queries were employed to create one combined CSV dataset, with input and engineered features from both databases. The features included are primarily about the total number of reviews (*SUMTotalReviewsOnSite*) or the positioning of the hotel rating (*AVGNormalizedRating*) concerning the competitor set's ratings (*AVGCompSetNormalizedRating*, *MedianCompSetAVGNormalizedRating*).

The merging of both databases into one dataset required some preprocessing. Although online reviews on almost every website have a similar structure (Bjørkelund, Burnett, Nørvag 2012), some important differences should be taken into account. For example, while both Booking.com and Tripadvisor.com reviews feature an overall rating and a textual component, Booking.com's rating is in a continuous range from 1 to 10, but Tripadvisor.com's uses a discrete range from 1 to 5. There is also a major difference in the textual component: Booking.com provides two text fields, one for positive and one for negative comments, while Tripadvisor.com only uses one single text field. Another important difference is how the two websites present ratings: although both sources allow users to assign ratings by concepts (cleanliness, location, comfort, among others), Booking.com presents aggregated results per hotel, while Tripadvisor.com presents results by review. Metadata and segmentation information of the reviews, such as age group, travel reason, or country of the reviewer, could be of importance, but in most social media websites it is not mandatory for reviewers to fully identify themselves, allowing to maintain anonymity (European Commission 2014). Therefore, even though segmentation information could be captured in some of the reviews because it was not available in all, this data was not considered of quality and, consequently, discarded. Due to the differences in the ratings scales of both websites and to the Booking.com rating scale distortion, which in fact has a minimum rating of 2.5 and not of 1 (Mellinas, María-Dolores, García 2016), ratings were normalized to a value ranging from 1 to 100. Normalization was done by using one of the most common normalization methods to scale variables, the min-max formula (Abbott 2014):

$$x' = \frac{(x - \min(x))}{(\max(x) - \min(x))} \times 100 \qquad (1)$$

This scale is typically used for indexes that aggregate ratings from multiple sources, such as the one used by Anderson (2012).

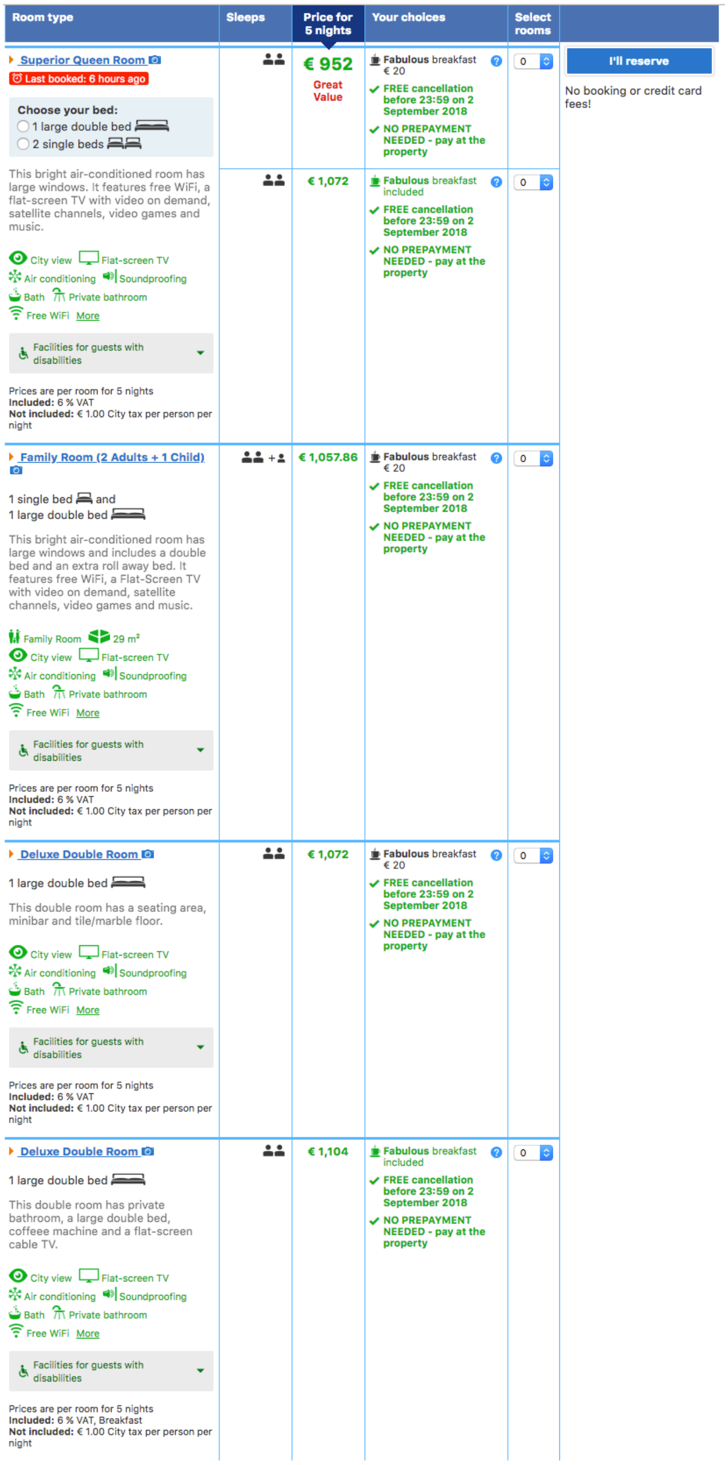
4.2.2.1.3 Online prices/inventory data

Research shows that in most hotel markets demand is relatively inelastic. Still, within a market, when in similar circumstances, a price lower than the competitive set price will drive share to the hotel (Enz, Canina, Lomanno 2009; McGuire 2016). For this reason, in the context of bookings cancellation prediction, online prices and inventory availability data is necessary to understand whether or not the fact that charging a customer with a higher price than its competitors may lead "deal-seeking" customers to cancel their booking.

Booking.com was selected as the source for acquiring the data due to its predominance in Europe (HOTREC - Association of Hotels, Restaurants and Cafes and similar establishments of Europe 2016; Martin-Fuentes, Mellinas 2018). Booking.com is so dominant that it can impose rules to hotels, such as not allowing hotels to close sales in Booking.com if they have sales open for other channels. To circumvent this imposition, when in situations where hotels intend to sell the rooms they have available in other distribution channels (e.g. in their own website), but not in Booking.com, hotels do so by raising prices in Booking.com to extreme highs. Due to this situation, the analysis of the positioning of a hotel against it competitive set is not usually done with the mean of the competitors' prices, but with the median, since the mean has the problem of being distorted by extreme values at either side of the distribution (Enz, Canina, Walsh 2001). Considering this, the competitive set's price feature constructed from the collected data, reflected the competitive set median price and (*MedianCompSetPrice*). The other features included in the dataset were essentially about prices, such as the hotel minimum hotel price *(MinPrice*), or the competitive set's minimum price (*MinCompSetPrice*), or about the number of rooms the hotels and their competitive sets had on sale (*CompSetMaxAvailableRooms*, *HotelsWithOpenSales*, *MaxAvailableRooms*). Please note, however, that if there are more than ten rooms on sale for a specific rate, Booking.com shows ten as the quantity available. Again, SQL queries were used to build the dataset in the CSV format.

One of the major differences between the extraction of online prices and inventory data from other data is the bi-dimensionality within time. Because hotel prices can change on a daily basis (or sometimes even more frequently) (McGuire 2016), in the context of cancellations, to explore the influence in cancellations that the relationship between the price a customer has agreed to pay to a hotel he/she made a booking and the prices at which the hotel competitors are selling a similar "product", data has to take in account two-time dimensions. The date when prices are being compared at (observation date) and the dates for the booked staying period (lookup date). This situation means that, for assessing the effects of prices on cancellations, prices and inventory information had to be collected every day (observation date), for the following 365 days (lookup dates). In addition, there is the question of the multiple combinations of room types and meal types each hotel offers for a determined number of persons. An example of this multiplicity is displayed in Figure 32, where just for a two-person occupation, one hotel shows 5 different rates, according to the room type and meal included.

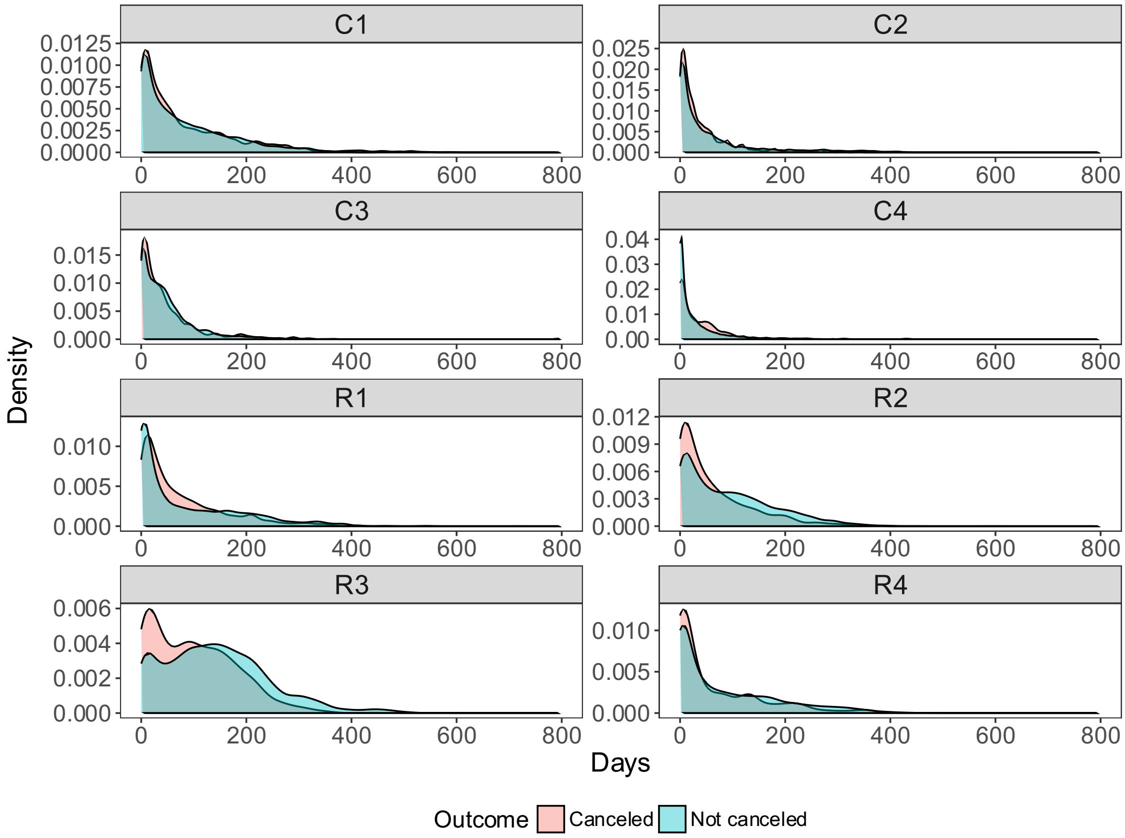
Figure 32 - Booking.com room type selection form



Room type	Sleeps	Price for 5 nights	Your choices	Select rooms	
Superior Queen Room Last booked: 6 hours ago **Choose your bed:** ○ 1 large double bed ○ 2 single beds This bright air-conditioned room has large windows. It features free WiFi, a flat-screen TV with video on demand, satellite channels, video games and music. City view Flat-screen TV Air conditioning Soundproofing Bath Private bathroom Free WiFi More Facilities for guests with disabilities Prices are per room for 5 nights **Included:** 6 % VAT **Not included:** € 1.00 City tax per person per night	👤👤	€ 952 Great Value	Fabulous breakfast € 20 ✓ FREE cancellation before 23:59 on 2 September 2018 ✓ NO PREPAYMENT NEEDED - pay at the property	0	I'll reserve No booking or credit card fees!
	👤👤	€ 1,072	Fabulous breakfast included ✓ FREE cancellation before 23:59 on 2 September 2018 ✓ NO PREPAYMENT NEEDED - pay at the property	0	
Family Room (2 Adults + 1 Child) 1 single bed and 1 large double bed This bright air-conditioned room has large windows and includes a double bed and an extra roll away bed. It features free WiFi, a Flat-Screen TV with video on demand, satellite channels, video games and music. Family Room 29 m² City view Flat-screen TV Air conditioning Soundproofing Bath Private bathroom Free WiFi More Facilities for guests with disabilities Prices are per room for 5 nights **Included:** 6 % VAT **Not included:** € 1.00 City tax per person per night	👤👤 + 👤	€ 1,057.86	Fabulous breakfast € 20 ✓ FREE cancellation before 23:59 on 2 September 2018 ✓ NO PREPAYMENT NEEDED - pay at the property	0	
Deluxe Double Room 1 large double bed This double room has a seating area, minibar and tile/marble floor. City view Flat-screen TV Air conditioning Soundproofing Bath Private bathroom Free WiFi More Facilities for guests with disabilities Prices are per room for 5 nights **Included:** 6 % VAT **Not included:** € 1.00 City tax per person per night	👤👤	€ 1,072	Fabulous breakfast € 20 ✓ FREE cancellation before 23:59 on 2 September 2018 ✓ NO PREPAYMENT NEEDED - pay at the property	0	
Deluxe Double Room 1 large double bed This double room has private bathroom, a large double bed, coffeee machine and a flat-screen cable TV. City view Flat-screen TV Air conditioning Soundproofing Bath Private bathroom Free WiFi More Facilities for guests with disabilities Prices are per room for 5 nights **Included:** 6 % VAT, Breakfast **Not included:** € 1.00 City tax per person per night	👤👤	€ 1,104	Fabulous breakfast included ✓ FREE cancellation before 23:59 on 2 September 2018 ✓ NO PREPAYMENT NEEDED - pay at the property	0	

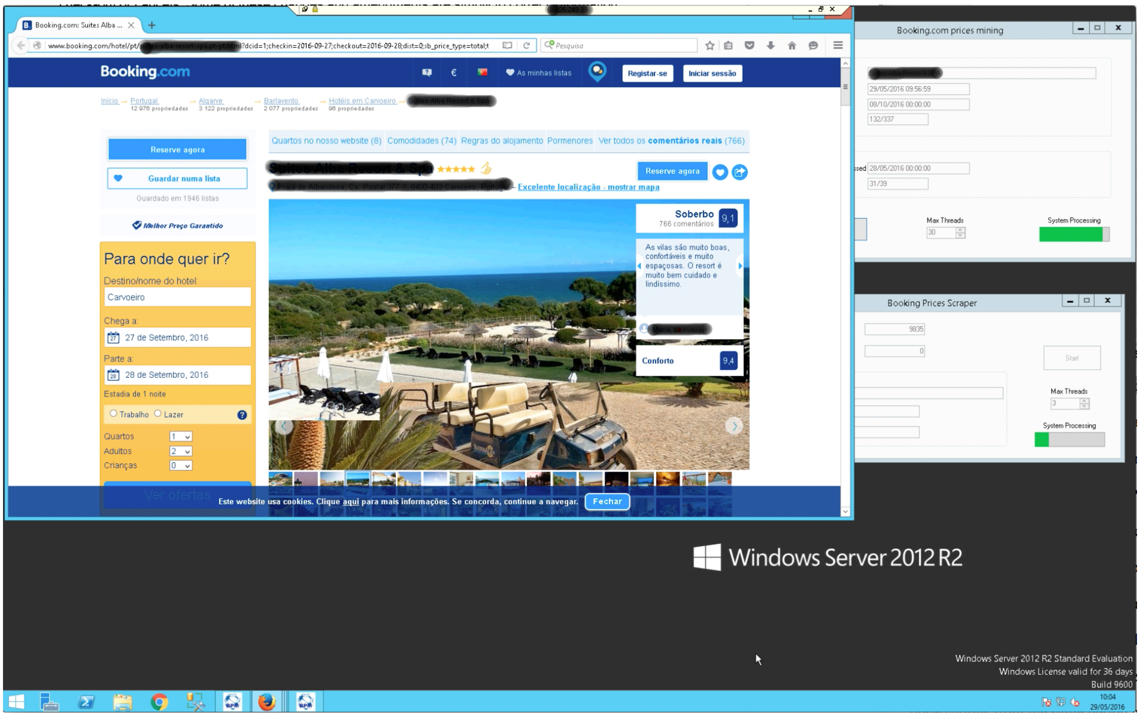
The volume of the information extracted, as described in Appendix E, is over 16 million observations of hotel, observation date, and lookup date combinations. This volume translated into almost 90 million observations by room type, maximum occupation, and meal combinations. The definition of the period of 365 days was based on the fact that the vast majority of bookings, independently of their cancellation outcome have a lead time far below that number of days (Figure 33). For bookings entered for a period out of the 365 following days, it was assumed that the competitors' price would be equal to the hotel.

Figure 33 – Lead time by cancellation outcome



As before with extraction of online reviews data from Booking.com, the extraction of prices and inventory on sale required the development of a custom-built extractor to scrap Booking.com web pages. Again, the extractor was built in C#, using Microsoft .Net Framework 4.5 and Selenium. SQL was used has the database to store the extracted data and metadata. Diagrams and dictionaries with detailed information of the database structure, including all metadata and data fields stored, as well as simple statistics of the data collected are available in Appendix E. However, due to the volume of data to be captured and processed, a different architecture had to be used. In this case, as illustrated in Figure 34, the extractor was divided into two components: one component for downloading the content of the prices/inventory web page (main component), and another to process the downloaded content and identify prices per room type, meal types, and maximum occupation (scraper component). Both components had multithread capability to take advantage of parallel processing. This extractor architecture allowed the scraper component to be deployed in multiple computers to take advantage of distributed computing, so that it would be possible to daily check the prices and inventory of all studied hotels and their competitors, for the following 365 days. Even running on three virtual Windows 2012 servers, with 32 Gb of RAM and 16 virtual CPU's each, it took the extractor between 12 to 16 hours to collect all the required data, every day.

Figure 34 - Screenshot of both components of Booking.com prices/inventory extractor



www.booking.com/hotel/pt/[redacted]?dcid=1;checkin=2016-09-27;checkout=2016-09-28;dist=0;sb_price_type=total;t

Início → Portugal 12 976 propriedades → Algarve 3 122 propriedades → Barlavento 2 077 propriedades → Hotéis em Carvoeiro 96 propriedades → [redacted]

Guardado em 1946 listas

Para onde quer ir?

Destino/nome do hotel: Carvoeiro

Chega a: 27 de Setembro, 2016

Parte a: 28 de Setembro, 2016

Estadia de 1 noite

○ Trabalho ○ Lazer

Quartos 1

Adultos 2

Crianças 0

[redacted] ★★★★★

[redacted] – Excelente localização - mostrar mapa

Soberbo 9,1 766 comentários

As vilas são muito boas, confortáveis e muito espaçosas. O resort é muito bem cuidado e lindíssimo.

[redacted]

Conforto 9,4

Este website usa cookies. Clique aqui para mais informações. Se concorda, continue a navegar. Fechar

Booking.com prices mining

[redacted]

29/05/2016 09:56:59

08/10/2016 00:00:00

132/337

ssed 28/05/2016 00:00:00

31/39

Max Threads 30

Booking Prices Scraper

9835

0

Max Threads 3

Windows Server 2012 R2 Standard Evaluation
Windows License valid for 36 days
Build 9600

10:04
29/05/2016

In spite of the efforts and resources put into the development, deployment and monitoring of the online prices and inventory extractor, due to difficulties associated to Booking.com constant changes on prices web pages' structure and the persistent running of A/B tests, only by mid-July 2016 was possible to assure the quality of the extracted data. In light of this situation, it was decided to use data from this source collected only after July 2016.

4.2.2.1.4 Additional data sources

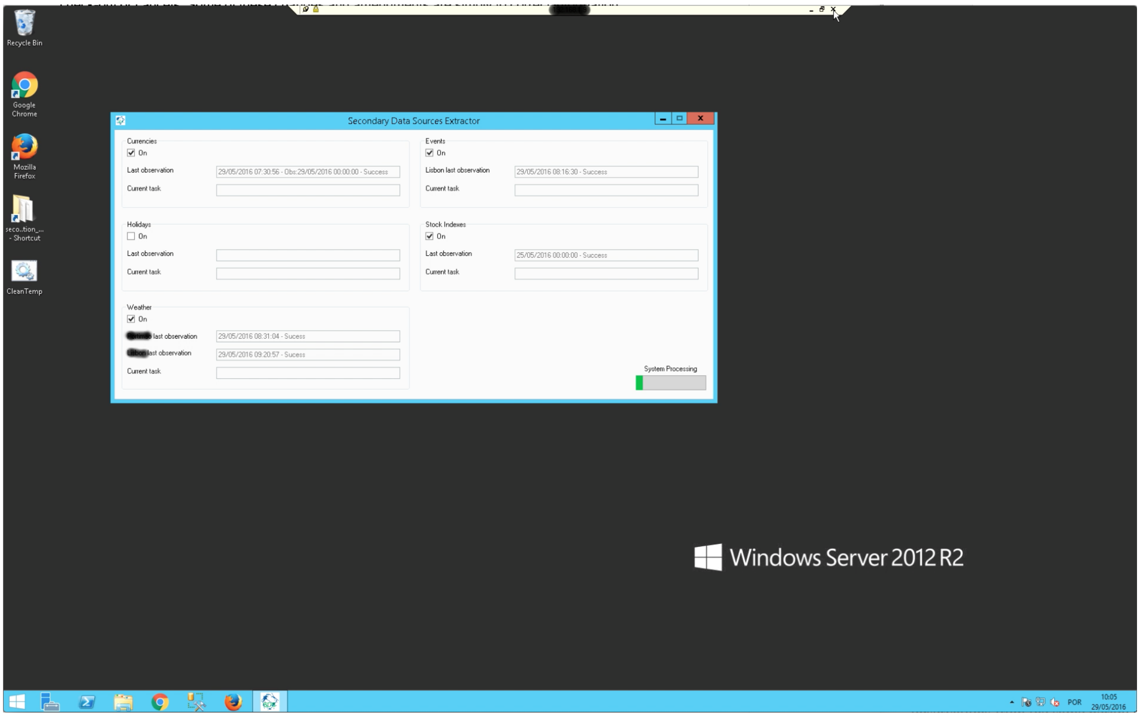
The extraction of data from the remaining sources, namely: holiday calendars, local weather forecasts, special events calendars, currency exchange rates, and stock exchange indexes required fewer resources and generated fewer challenges than the previous extractions. First, because for some sources, such as currency exchange rates, holidays calendars, and weather forecast, data could be extracted via APIs instead of web scraping, which decreased the possibility for the occurrence of problems during the extraction. Second, because even for sources were data needed to be scraped (special events calendars and stocks exchange indexes), the websites did not employ any special tactics to difficult web scraping, nor were they dynamic in the sense of regularly changing their content or structure. Third, because the volume of data associated to each source was relatively small.

These type of data sources were selected as a way to try to understand the influence of the factors associated to each data source in the cancellation of bookings. For example, in terms of demand, it is known that precipitation should be considered more important than temperature when pondering weather impacts (Day, Chin, Sydnor, Cherkauer 2013). Therefore, a feature that captures raining probability for futures dates could, presumably, help explain cancellations of a certain type of customers (e.g., "city-break" or "long-weekend" customers in opposition to corporate customers). If conjugated with the information about the staying period of the customer including a holiday or a special event in the region, features for these sources may help predict cancellations. Additionally, features derived from the fluctuation of currencies exchange rates and stocks exchange indexes between the country of the booking origin and Portugal could help explain if changes in macroeconomic conditions in some countries of origin could impact cancellations. However, during data understanding and data preparation phases, it was confirmed that the country of origin field in bookings is only correctly entered and verified at check-in. Therefore, using currencies exchange rates and stocks exchange indexes for incorrect countries did not make sense in practical terms. As such, no datasets were created for these sources of data. Datasets were indeed created for holiday's calendar, special events and weather forecast sources. Even so, the holiday's calendar dataset creation was also influenced by the country field aforementioned condition. Because of this, it was decided to include only Portuguese holidays in the calendar. The rationale behind has to do with the fact that the vast majority of customers who stay at Portuguese hotels share a high number of holidays with Portugal (Instituto Nacional de Estatística 2016).

Once again, to reduce storing and processing requirements, the creation of these sources' datasets involved the creation of new features and some preprocessing. Mainly, the holidays calendar dataset included features to describe the date of the holiday and its designation. The special events dataset included features about the location of the event (resort or city), type and date. The weather forecast included features related to current conditions and conditions forecasted for the following ten days, including temperature, wind and rain probability and quantity.

This extractor was also built in C#, using Microsoft .Net Framework 4.5 and Selenium. As shown in Figure 35, the extractor was deployed in a virtual Windows 2012 server, with 8 Gb of RAM, and 2 virtual CPU's. SQL was used as the database to store the data extracted from these five data sources. Diagrams and dictionaries with detailed information of the database structure, including all metadata and data fields stored, as well as simple statistics of the data collected, are available in Appendix E. In average, it took the extractor, daily, two hours to extract data from all the five sources.

Figure 35 - Screenshot of the additional data sources extractor



Secondary Data Sources Extractor

Currencies
☑ On
Last observation: 29/05/2016 07:30:56 - Obs:29/05/2016 00:00:00 - Success
Current task:

Events
☑ On
Lisbon last observation: 29/05/2016 08:16:30 - Success
Current task:

Holidays
□ On
Last observation:
Current task:

Stock Indexes
☑ On
Last observation: 25/05/2016 00:00:00 - Success
Current task:

Weather
☑ On
[redacted] last observation: 29/05/2016 08:31:04 - Sucess
[redacted] last observation: 29/05/2016 09:20:57 - Sucess
Current task:

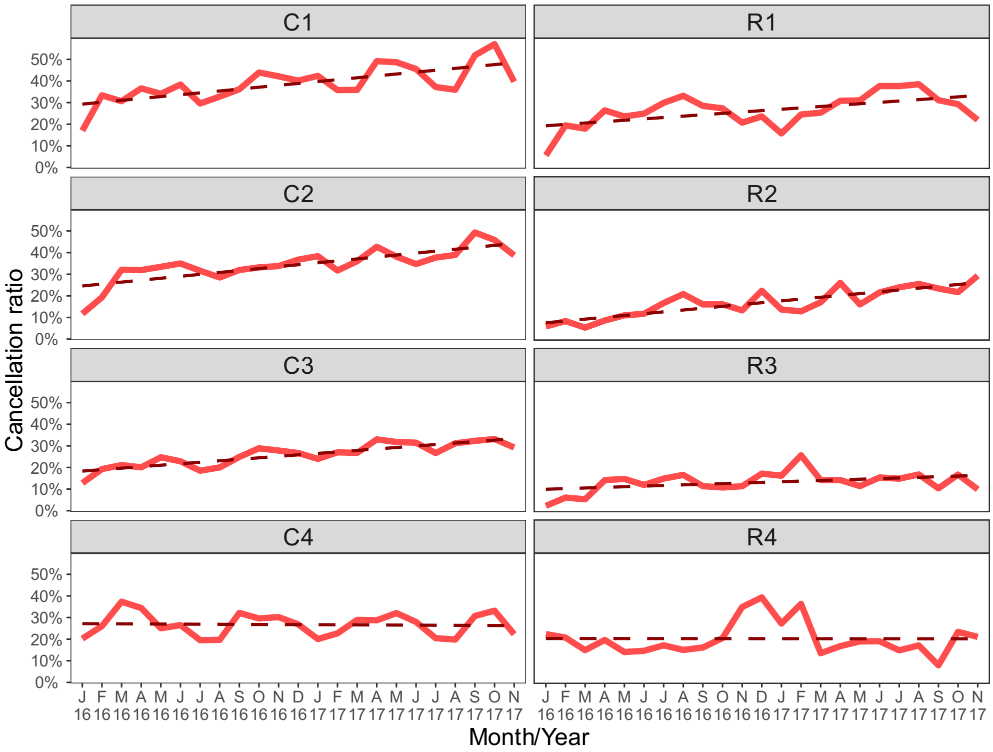
System Processing

4.2.2.2 Data quality, description, and exploration

Once again, summary statistics and data visualization were used to understand final models' data. The summary statistics for the final models' PMS data showed similar abnormalities to the ones previously observed in PMS data employed in the exploratory models. In the versions of C1 and R1 datasets with additional features, missing values were found for the features: *Agent, Country*, and *Company*. The missing values can be explained by the way datasets were created. Sometimes, fields not fulfilled at booking are left empty instead of displaying a "NULL" value. In terms of numeric/integer features, most data quality situations are related to the existence of outliers or high positive/negative skew. These situations are explained not by errors in data extraction or data entering, but by normal hotel operations, such as canceled group bookings, corrections or amendments, abnormally high demand days (e.g. Pope visit). Examples of such situations are visible in almost all PMS' datasets, in features like *ADR*, *Adults*, BookingChanges, *Children*, *LeadTime* (clearly visualized in Figure 33), PreviousBookingsNotCanceled, *PreviousCancellations*, StaysInWeekendNights, *StaysInWeekNights*, among others. Apart from these issues no major missing values, outliers, skewness, or other types of problems were identified, which denotes the overall good quality of PMS data. The same also applied to the other data sources. The only situation worth mentioning was the weather forecast data that, for city hotels, is missing during 8 days.

Besides showing patterns and tendencies in PMS data similar to the ones found for the exploratory models, data exploration showed that, in the new period of study, cancellations were increasing for almost all of the hotels, as illustrated by the trend lines in Figure 36. An analysis by hotel type and year, month, week and weekday, emphasizes this tendency. With Figure 37 it is possible to confirm that cancellation ratios in 2017 were higher than in 2016 for both hotel types. The figure also shows that cancellation rates for city hotels tend to be higher than for resort hotels. Other than that, no significant patterns seem to exist per month, week or weekday. However, on a closer look, by arrival date weekday, some patterns seem to exist (Figure 38). For resort hotels, cancellations ratio in Sundays is usually higher than in other weekdays. The same does not apply to city hotels. In the case of Tuesdays and Wednesdays, the patterns seem to be similar for both hotel types. These patterns point out that the arrival date weekday could have higher predictive power (Morales, Wang 2010).

Figure 36 - Cancellation ratio evolution



PMS data visualization shows some differences between the hotels in terms of cancellations by customer type, repeated guests and deposit types (Figure 39). Resort hotels and C1 work more with groups and contracts than others. R3 mainly works with groups and contracts, having very few transient and transient-party customers. Except for R4 and C1, hotels seem to have cancellation policies allowing customers to be refunded in case of cancellation. C1 might have exceptions for refunding, as most of the bookings in that condition were canceled. Figure 39 also displays some interesting cancellation patterns, particularly for transient and transient-party customers with non-refundable cancellation policies/deposit types. Contrary to what was to be expected, this type of policy presents relatively more cancellations. Further analysis by country of origin, distribution channel and agent confirm what has been recognized by hotel managers: that most of the canceled bookings were not made with the intention to book a room, but with the purpose of having a proof of reservation. In fact, a hotel booking is mandatory for applying for a Portuguese entry visa. These bookings usually came through OTAs and presented false or invalid credit card details. The hotel identifies these bookings as “fake” after failing to charge the credit card, and contact the customer. However, during the time required to verify the credit card, the bookings contribute negatively to demand forecast and demand management decisions.

Figure 37 - Cancellation ratio by hotel type and time dimensions

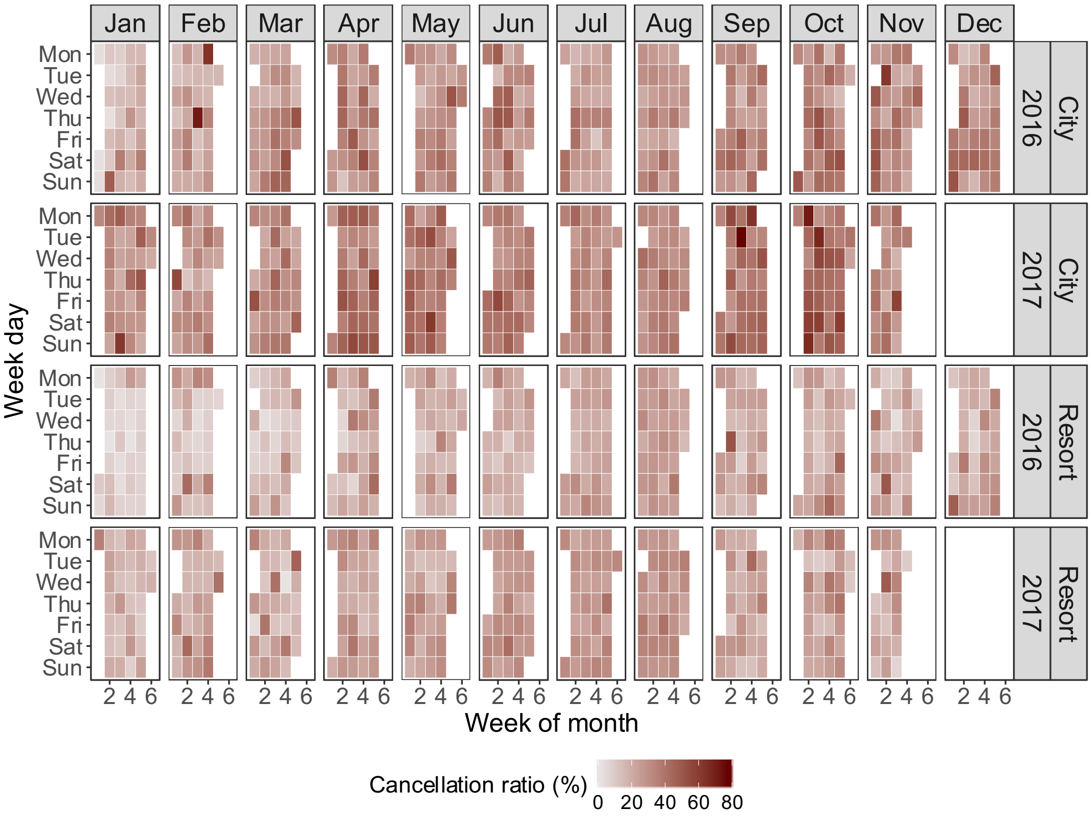


Figure 38 - Cancellation ratio by weekday

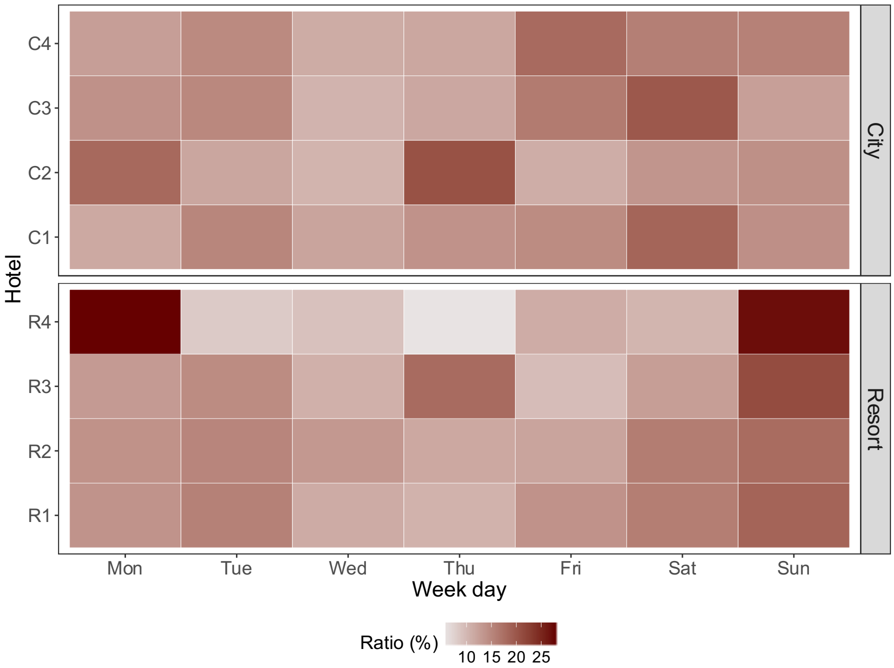
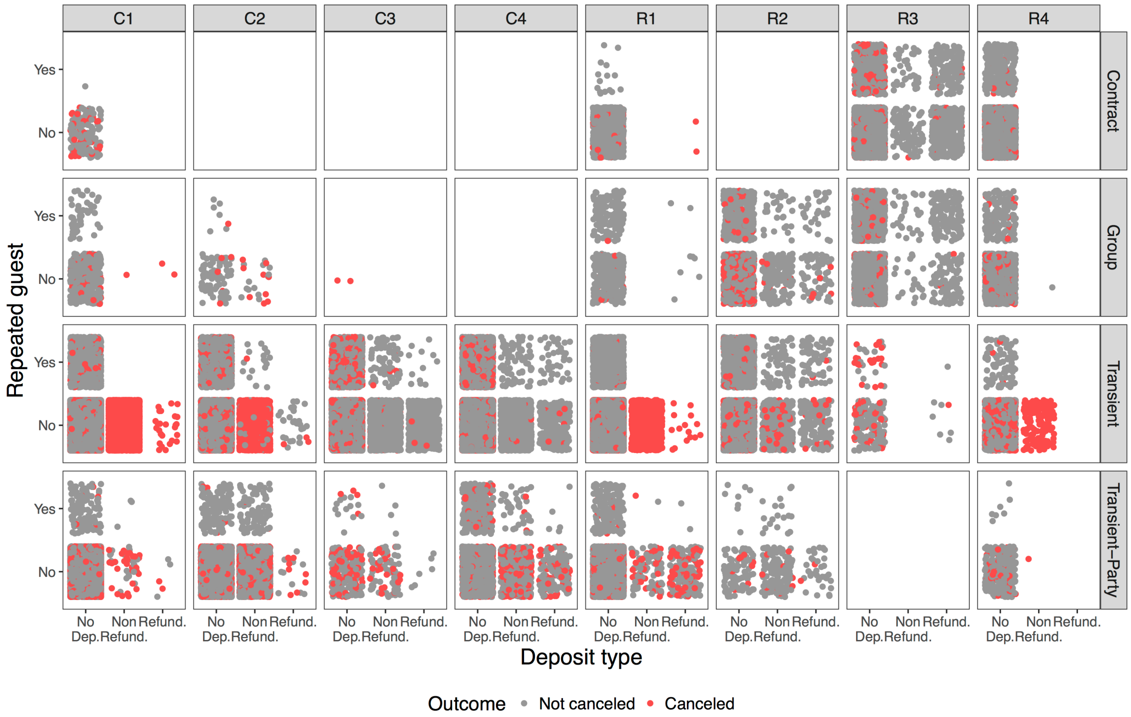


Figure 39 - Cancellations by deposit type, guest type, and repeated guest



The combination of PMS data with data from other sources reveals new angles in terms of explaining cancellations. Starting by the relation between holidays and weather forecast, chi-square tests of independence show that, for all city type hotels and for R1, there is a significant association between cancellations and bookings ($p<0.05$) when the period of stay covers a holiday (Table 6). A post hoc analysis of the residuals shows there are more cancellations than expected in bookings covering a holiday than in bookings where the period of stay does not, the pattern that is visible in Figure 40.

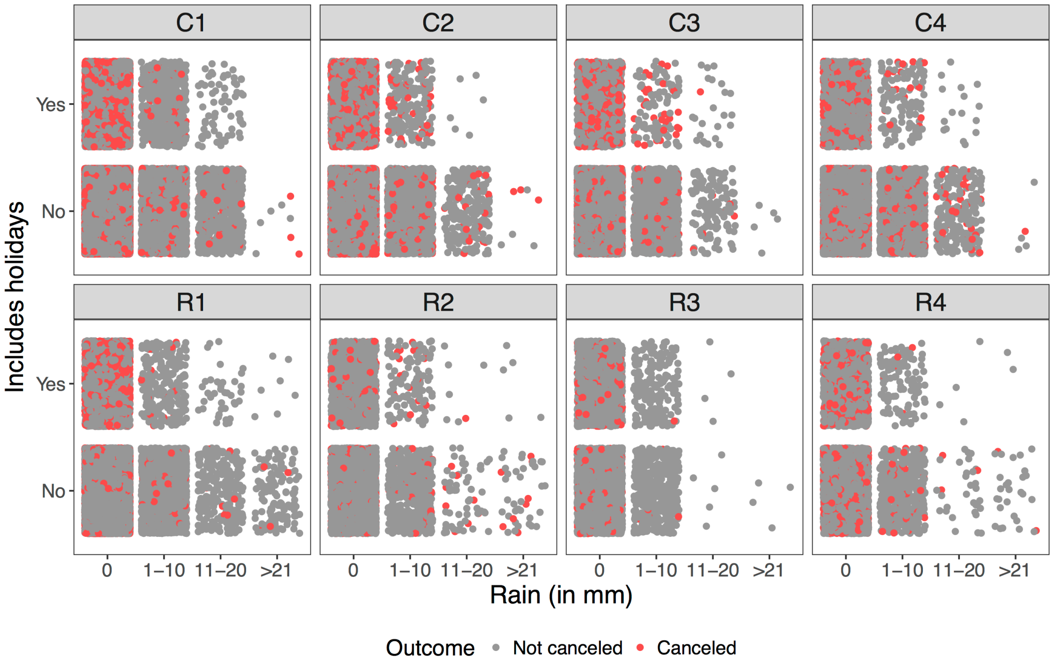
There is also a significant association between cancellations and rain forecasting, for every hotel. However, in this case, post hoc analysis shows the relationship is inverse, that is, there are fewer cancellations than expected when rain is forecasted. This can be explained by the fact that when bookings are canceled with 10 or more days of the expected arrival date, the feature's value for *AvgQuantityOfPrecipitationInMM* is calculated as 0 for the days outside of the 10-day window for which weather forecasts exist.

Table 6 - Chi-square test results between cancellations outcome, holidays, and rain forecast

Hotel	Variable	X-squared	Degrees of freedom	p-value
C1	Include holidays	7.0582	1	0.0079
	Rain forecast	2075.9	3	<0.0001
C2	Include holidays	10.936	1	0.0009
	Rain forecast	579.13	3	<0.0001

Hotel	Variable	X-squared	Degrees of freedom	p-value
C3	Include holidays	11.718	1	0.0006
	Rain forecast	192.9	3	<0.0001
C4	Include holidays	9.2105	1	0.0024
	Rain forecast	238.49	3	<0.0001
R1	Include holidays	12.549	1	0.0004
	Rain forecast	771.82	3	<0.0001
R2	Include holidays	1.8914	1	0.1690
	Rain forecast	36.93	3	<0.0001
R3	Include holidays	3.1627	1	0.0753
	Rain forecast	76.887	3	<0.0001
R4	Include holidays	0.2045	1	0.6511
	Rain forecast	66.196	3	<0.0001

Figure 40 - Cancellation outcome according to weather forecast and holidays during stay

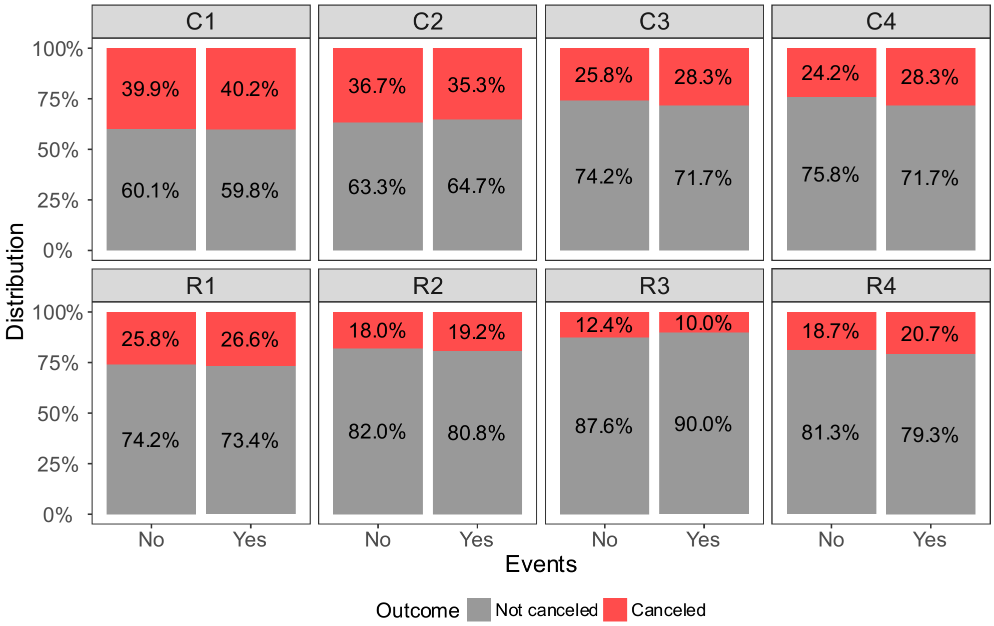


Regarding the impact that events may have on cancellations, chi-square test results are mixed (Table 7). For hotels C2, C3, C4, and R3, cancellation, although presenting different patterns, is significantly related to the existence of events during the period of stay (p<0.05). The same does not happen for C1, R1, R2, and R4. Notice also, that, for hotels C3 and C4, the number of cancellations is higher when there is an event, while the opposite happens in C2 and R3 (Figure 41).

Table 7 - Chi-square test results between cancellations outcome and events in hotels' region

Hotel	X-squared	Degrees of freedom	p-value
C1	0.4383	1	0.5080
C2	4.3993	1	0.0360
C3	7.1489	1	0.0075
C4	35.2230	1	<0.0001
R1	0.6998	1	0.4029
R2	0.7363	1	0.3909
R3	4.5869	1	0.0322
R4	1.9903	1	0.1583

Figure 41- Cancellation outcome for days of events

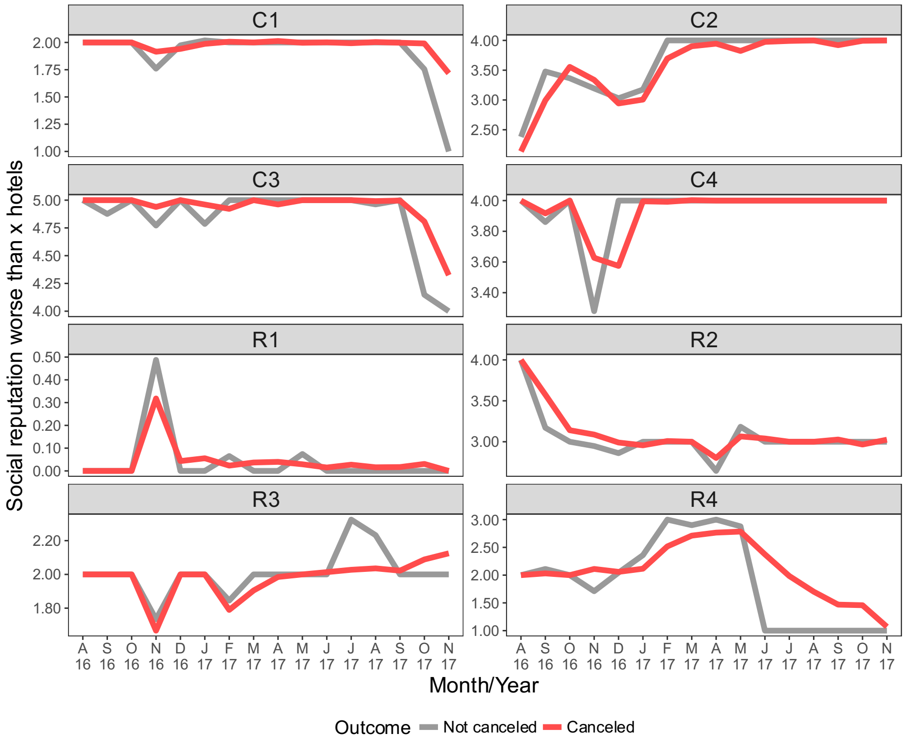


Regarding social reputation, different reputation dimensions were evaluated (such as the reviews variance and volume per hotel) but the global rating was the one used since it showed the higher explanatory power, as already identified by Viglia, Minazzi, Buhalis (2016). When comparing the number of hotels from the competitive set presenting better social reputation rating at arrival date (for non-canceled bookings) or cancellation date (for canceled bookings), fuzzy results are obtained. An analysis of the *WorseThan* feature by booking cancellation outcome (*IsCanceled*) shows that only for hotels C1, C2, C3, R3, and R4, the relation was significant (p<0.05) (Table 8). The differences can be seen in Figure 42, which shows the monthly average of the two features by cancellation outcome.

Table 8 - Kruskal-Wallis chi-squared results between cancellations outcome and competitive set social reputation positioning

Hotel	X-squared	Degrees of freedom	p-value
C1	561.26	1	<0.0001
C2	34.735	1	<0.0001
C3	83.049	1	<0.0001
C4	2.3616	1	0.1244
R1	1.7015	1	0.1921
R2	1.96	1	0.1615
R3	8.1743	1	0.0042
R4	58.882	1	<0.0001

Figure 42 – Monthly average competitive set social reputation positioning at cancellation outcome date

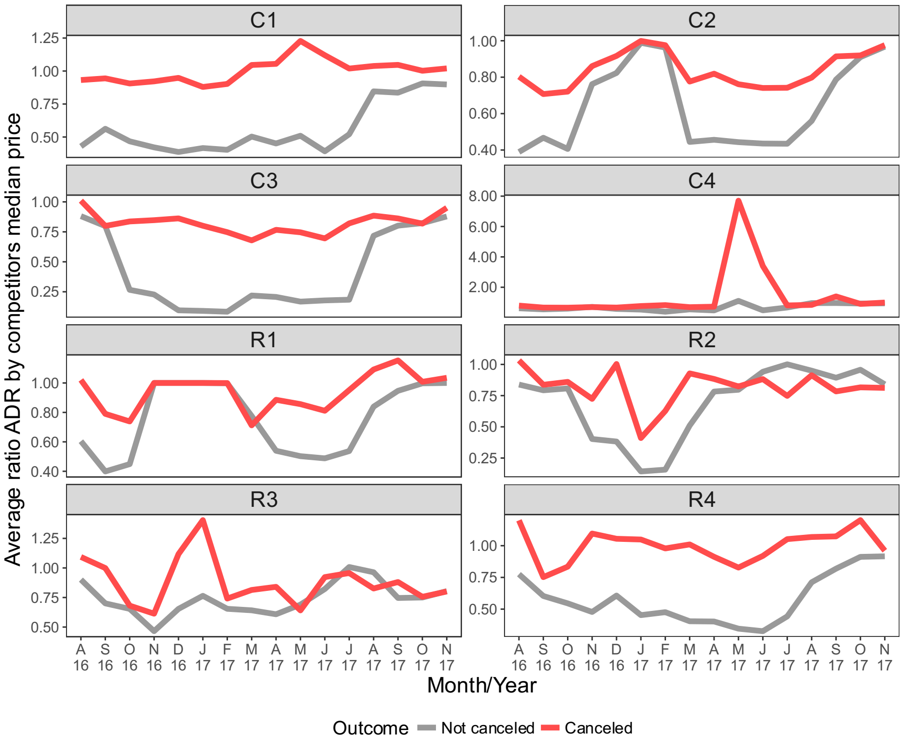


Competitors' prices seem highly related to cancellations. The analysis of the feature *RatioADRbyCompsetMedianDifference*, calculated by dividing the booking ADR by the competitors set median price for the booking period of stay and at the time of cancellation outcome (arrival date or cancellation date) shows this relation is statistically significative (p<0.05) for 7 of the 8 hotels (Table 9). This relation is clearly visible in Figure 43, where it is possible to see that mostly the ratio for canceled bookings is superior to the non-canceled. The only hotels for which this difference is not so clearly visible are R2 and R3, which, are the two hotels where OTA's have a lesser expressive distribution share, as depicted earlier in Figure 31. This OTAs expression may help to explain the results presented in Table 9, in the sense that, because these hotels do not rely so much on OTAs, they are not so exposed to the so-called "deal-seeking" customers.

Table 9 - Kruskal-Wallis chi-squared results between cancellations outcome and a ratio of the hotel ADR by competitors' set median price

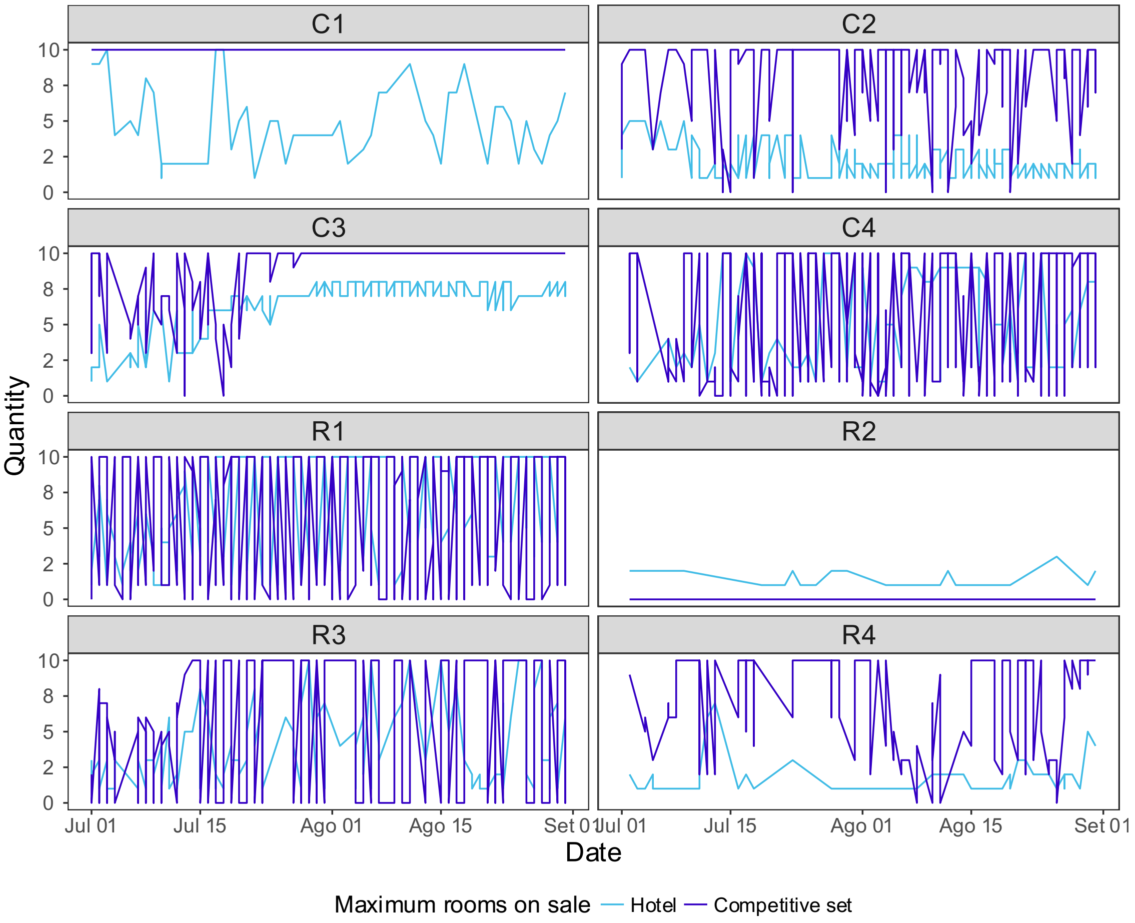
Hotel	X-squared	Degrees of freedom	p-value
C1	10663	1	<0.0001
C2	1083.1	1	<0.0001
C3	937.8	1	<0.0001
C4	49.717	1	<0.0001
R1	463.04	1	<0.0001
R2	1.8232	1	0.1769
R3	11.567	1	0.0007
R4	655.28	1	<0.0001

Figure 43 - Monthly average ratio of ADR by competitive set median price at cancellation outcome date



Due to the two-dimensional nature of the inventory for sale data, analysis of patterns have to be made by observation date for specific lookup dates. In this case, no striking patterns emerged from the analysis. For example, for the observation date July 1st, 2017, for the lookup period starting ending in August 31st, 2017, no unusual pattern is perceived (Figure 44). The sole point out is the low number of rooms that R2 and its competitors put on sale on Booking.com for most of the year. For what was understood, this seems to be related to the fact that the hotels' distribution is mostly assured by traditional tour operators, which was previously recognized and acknowledged by the hotel's manager.

Figure 44 - Inventory on sale on the 1st July 2017 for the following two months



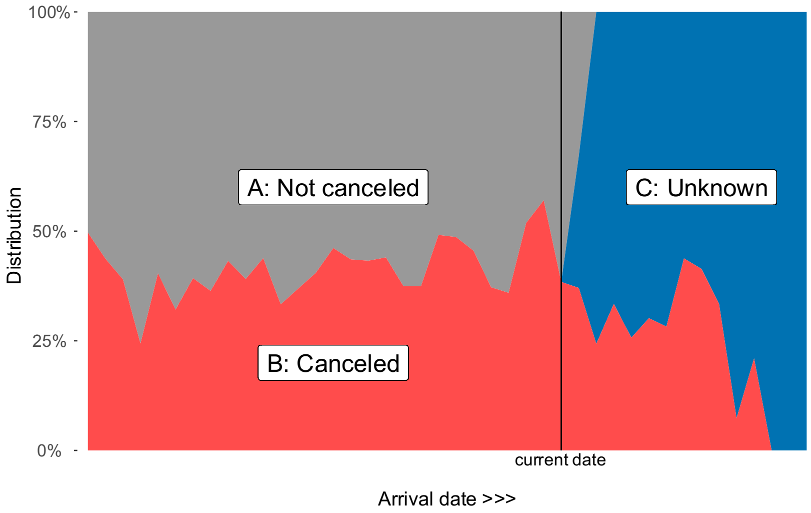
Although exploratory models yield good results, they also revealed a tendency for overfitting data. Consequently, models did not generalize well for unknown bookings - a common issue in machine learning models (Domingos 2012). Data understanding revealed two issues that probably had a considerable influence in the performance of the models: data leakage and "dataset shift", i.e., *"where the joint distribution of inputs and outputs differs between training and test stage"* (Quiñonero-Candela, Sugiyama, Schwaighofer, Lawrence 2009, p. xi). This distribution shift occurred for two reasons. First, due to the speed at which the hospitality business changes. The stratified dataset splitting strategy for the creation of the training and testing datasets does not guarantee a comparable distribution among both datasets. Second, the rapid growth of the tourism industry in recent years and the increasing annual demand causes a rapid increase in the prices *ADR* and *LeadTime*, which contribute to differences in the distributions of inputs and outputs over time. Also, this fast pace of operations causes the continued arrival of new players (OTAs) and the disappearance of other players, namely “traditional” travel agencies and travel operators. These constant transformations contribute to a change in the representative weight of these entities in the hotel operation, which influences the distribution of certain features over time such as *ADR*, *LeadTime*, *Agency* or *Company* (known as “concept drift”). To solve these issues, two major refinements were introduced into the models: changes in the dataset construction and dataset splitting, and changes in feature selection and engineering. These changes will be addressed in the following sections.

4.2.3 Data preparation

As expected, the starting point for building each hotel modeling dataset was the knowledge obtained during the development of the exploratory models, especially the identification of which PMS’s features should be used. This knowledge was complemented with the quality issues, and insights found during data understanding, which allowed for the selection of features, the creation of new features, cleaning and formatting of data, and lastly, to integrate the different data sources to build a unique dataset per hotel.

In terms of features exclusively built from PMS data, in comparison to the exploratory models, five general changes have been performed. First, the features *ArrivalDateDayOfMonth*, *ArrivalDateMonth*, *ArrivalDateWeekNumber*, and *ArrivalDateYear* were replaced by *DayOfYear*, which represents the sequential number of the day in the year (from 1 to 365/6). With this replacement, seasonality could still be captured by one only feature instead of four. At the same time, this also removed a leakage problem caused by *ArrivalDateYear*, whose use caused the models to learn that bookings for future years tend to be canceled. This situation is easily understandable by observation of Figure 45. Bookings with arrival for a future date cannot be used in the modeling dataset as their outcome is unknown (C: type bookings). Removal of this type of bookings makes future dates to be highly imbalanced since most bookings with a known outcome are canceled (B: type bookings). Second, to reduce the problem of how the distribution of the *ADR* changes over time, it was replaced by *ThirdQuartileDeviationADR*. The new feature is calculated by the dividing the *ADR* by the third quartile value of all bookings from the same distribution channel, same reserved room type and arriving at the same week. The feature was created to reflect how much was a customer paying for the same type of room when compared with other “similar” customers. By turning it to be a ratio would prevent it to reflect high demand peaks and make it more robust to outliers. Third, because not all of the hotels had parking spaces or use the PMS to control the access to parking spaces, the feature *RequiredCarParkingSpaces* was only included in Model 4 (the one that was built with additional features) and only for hotels C1 and R1. Fourth, still in Model 4, new features such as *AssociatedToEvent*, *BookedSPA*, *SRDoubleBed*, *SRHighFloor*, *SRQuitetRoom*, *SRTogether* and *SRTwinBed* were included to understand to what extent the effect of additional features could reflect hotels’ operations and services and improve the model's performance. Essentially, these new features indicated if the booking was associated to an event taking place in the hotel itself, if the customer has booked SPA treatments, or if the customer have made special requests (a type of bed, high floor, quiet room, among others), respectively. Fifth, new features *FolioNumber*, *ReservationStatus*, and *ReservationStatusDate* were also used but because of the integration and modeling process, not as modeling features.

Figure 45 - Bookings distribution by cancellation outcome (hotel C1 example)



The major challenge in terms of data preparation came from the non-PMS data sources, mainly because of computing and time resources required to engineer features, build models and evaluate them in order to decide which features would be included in the final models. After several iterations between the different process model phases, a set of new features was created to capture factors that could have predictive power. These were *AvgQuantityOfPrecipitationInMM,* to capture the average forecasted precipitation for the period of stay, *WorseThan,* to capture the positioning of the hotel in the booking outcome date (cancellation or arrival date), *HotelsWithRoomsAvailable,* to capture how many competing hotels had rooms on sale for the period of stay at the booking's outcome date, *nHolidays,* to capture if the period of stay covered holidays, *RatioMajorEventsNights* and *RatioMinorEventsNights* to capture the existence of events during the period of stay, and *RatioADRbyCompsetMediaDifference* to capture the deviation between the booking price and the median of the competitors' prices, at the booking's outcome date. More details on these features can be found in Appendix C.

Before the integration of all data sources, minor cleaning and formation operations were performed, namely removing observations who presented an *ADR* below zero, assigning the value "0" to the features *Agent* and *Company* who presented missing values, and drop non-used levels in categorical levels (these non-used levels appeared with the removing of the observations with an *ADR* below zero). Additionally, the missing values in weather forecast data were processed. R package "MissForest" (Stekhoven 2013) was used to determine values for the missing days. This package employs a random forest machine learning algorithm to train a model on observed values to predict the missing values. Only then was data from the different sources integrated and merged into one modeling dataset per hotel.

One last operation was then performed in each hotel's dataset. R package "vtreat" (Mount, Zumel 2017) was used to reformat the categorical features. Categorical features with a high degree of cardinality can make model training slow and overfit data (Abbott 2014), and as seen before, models that overfit do not generalize well. To avoid this, all levels of categorical features with a minimum frequency of 0.02 were encoded into an indicator column (one-hot encoding[7]). However, not to lose information about the less common levels, a new numeric feature for each categorical feature was built. This feature's value represents the Bayesian change in the logit-odds from the mean distribution conditioned on the observed value of the original value. Vtreat adds a suffix to the feature name according to the type of feature: "_clean" for numeric features, "_catB" for features that represent the Bayesian change of categorical features, and "_lev_x.<level name>" for indicator features for categorical levels with a frequency greater than 0.02.

4.2.4 Modeling

As previously introduced, most high-performance machine learning algorithms are mostly a black box that generates highly complex prediction equations (Kuhn, Johnson 2013). Nonetheless, some outputs, such as those based on decision trees, are easier to understand by humans (Abbott 2014; Hastie, Tibshirani, Friedman 2001; Kuhn, Johnson 2013). Decision tree-based algorithms also have the advantage of automatically incorporating the treatment of outliers, handle missing data well, are not affected by feature skewness, inherently detect feature interactions, are non-parametric (making no distribution assumptions about features and the outcome variable), and have a built-in feature selection mechanism (Abbott 2014; Kuhn, Johnson 2013). However, decision tree algorithms also have weaknesses, like non-adaptability to slight changes in data and not generalize well. To overcome these weaknesses, some approaches employ ensemble methods, which, by combining multiple trees into one model, tend to have better performance (Hastie, Tibshirani, Friedman 2001; Kuhn, Johnson 2013). Therefore, and because for the exploratory models, decision tree-based algorithms (Boosted Decision Tree and Decision Forest) had already presented the best results, it was decided to build the final models with the

[7] "One-hot encoding" or the creation of "dummy variables" is a technique employed for numeric representation of categorical data. This technique involves the replacement of the categorical feature by as many features as the number of distinct category levels (Abbott 2014). For example, if the categorical feature "RoomType" had three categories (standard, deluxe and suite), this feature would be removed and replaced by three new features, one for each level. Then, a binary value of 0 or 1 would be assigned to each of these features, according to the original category level of the observation. For example, if "RoomType" for a particular booking was for a "standard", then the new "standard" feature will be assigned a 1, and a 0 would be assigned to the features "deluxe" and "suite".

award-winning, ensemble tree-based XGBoost machine learning algorithm (Chen, Guestrin 2016), a gradient boosting-based algorithm. Gradient boosting algorithms are usually faster than other methods in training models and allow the understanding of the importance of each feature in the prediction of the outcome (Hastie, Tibshirani, Friedman 2001).

The effectiveness of XGBoost, particularly in terms of controlling overfitting, is achieved by a set of parameters that enable fine-tuning of the model's complexity, including parameters to add randomness to make training more robust to noise. These parameters include the definition of the subsample of observations to use in each decision tree and the definition of the subsample of features to use per decision tree and per tree level.

For the estimation of model parameters, including the learning rate and boosting, a combination of two well-known techniques—grid-search and random-search—was employed (Bergstra, Bardenet, Bengio, Kégl 2011). Parameter values were selected from the model that presented a better error rate, from a total of 100 iterations of ten-fold cross-validations, over a maximum ensemble of 200 trees. In cross-validation, the parameter "early stop" was set to eight, which indicated that training was stopped after eight rounds of training set error improvements without a correspondent improvement of the test set error to avoid overfitting. For each iteration, parameters were randomly selected according to limits that were previously established during manual optimization experiments. The list of parameters and source code to select its values and the established limits are provided in Table 10. "colsample_bytree" indicates the subsample of features to use in the construction of each tree, in each boosting iteration. "eta" indicates the step size used to update overfitting. "gamma" is a parameter used to define the minimum split loss (the larger the value, the more conservative is the model). "lambda" it is the L2 regularization parameter. The higher the value, the more conservative is the model. "max_delta_step" is used to make the update step more conservative. "max_depth" indicates the maximum depth of trees. The higher the value, the higher is the tendency for models to overfit. "min_child_weight" is the minimum sum of instance weight needed in a child. The higher the value, the more conservative will the model be. More details of these parameters can be found in the XGBoost documentation (Chen, Guestrin 2016).

Table 10 - Models' estimation parameters selection source code

Parameter	R source code
colsample_bytree	runif(1, 0.4, 0.8)
eta	runif(1, 0.01, 0.3)
gamma	runif(1, 0, 0.2)
lambda	runif(1, 0, 0.5)
max_delta_step	sample(1:5, 1)
max_depth	sample(2:4, 1)
min_child_weight	sample(1:5, 1)

To select the best models and assess their performance, since this was a data-rich situation, the datasets were split according to the approach recommended by Hastie, Tibshirani, Friedman (2001) of dividing the datasets into three parts: a training set for fitting the model, a validation set for assessing the prediction error, and a test set to assess the generalization error. There is no specific rule to define the number and which observations are to be included in each of the sets since it depends on the characteristics of the data, such as size and structure (Guyon, Elisseeff 2003; Hastie, Tibshirani, Friedman 2001; Kuhn, Johnson 2013). However, as previously mentioned, for this type of problem time is not an irrelevant dimension. For example, the more cancellations a customer has made in the past, the higher the customer's probability to cancel. Consequently, this can be considered a temporal data problem and thus data for the test set should be chosen from a period not "known" by the training and validation set (Abbott 2014; Hastie, Tibshirani, Friedman 2001). Therefore, *ReservationStatusDate* was defined as the date to use for the splitting point in the creation of the test set. All bookings that were canceled or checked-in after August 31st, 2017, formed the test set. Considering the existing "dataset shift" problem, a method borrowed from time series techniques was employed to create the training and testing datasets: convenience splitting (Reitermanová 2010). This method enables the capture of "non-stationary temporal data": data that *"changes behavior with time and therefore should be reflected in the modeling data and sampling strategies"* (Abbott 2014, p. 197). Convenience splitting involves the division of the dataset in discrete "time" blocks. For this, bookings that were canceled or that had checked-in before August 31st, 2017, were divided into blocks of "month/year" arrival dates. From each block, 75% of bookings were assigned to the training dataset and the remaining 25% to the validation dataset.

Because data were not available for all data sources for the same period, it was decided to build different models using different datasets in terms of features and number of observations, as previously mentioned. The first model, using PMS features exclusively, encompassed arrivals from the January 1st, 2016 to November 20th, 2017 (Model 1). A second model, again using PMS features, used arrivals from August 1st, 2016 to November 20th, 2017 (Model 2). A third model (Model 3) included features from all sources with observations from the same period as Model 2. Lastly, a fourth model (Model 4) was created just for hotels R1 and C1, the ones that shared characteristics that allowed the creation of additional features. The period of observations for this model was the same as Models 2 and 3. The respective results are presented and discussed over the following section.

4.2.5 Evaluation

In this section, rather common machine learning metrics (described in Appendix A) are employed to present and discuss the results.

One of the first observations about the modeling results (Table 11) is that they differ, not only per model but within different hotels employing the same type of model.

Table 11 - Model's performance metrics

Hotel	**Model**	**Training set**			**Validation set**			**Test set**		
		Acc.	**Pre.**	**AUC**	**Acc.**	**Pre.**	**AUC**	**Acc.**	**Pre.**	**AUC**
C1	1	0.7844	0.7875	0.8767	0.7775	0.7838	0.8680	0.7755	0.7288	0.8636
	2	0.8050	0.7916	0.9007	0.7967	0.7778	0.8904	0.8323	0.7599	0.9226
	3	0.7887	0.7957	0.8799	0.7777	0.7769	0.8662	0.8122	0.7491	0.8964
	4	0.8350	0.8124	0.9242	0.8266	0.8033	0.9146	0.8490	0.7699	0.9319
C2	1	0.8294	0.7993	0.9103	0.8165	0.7786	0.9103	0.7686	0.5698	0.8271
	2	0.8493	0.8044	0.9307	0.8280	0.7790	0.9307	0.7863	0.5994	0.8474
	3	0.8385	0.8065	0.9183	0.8096	0.7673	0.9183	0.7851	0.5951	0.8422
C3	1	0.8497	0.7887	0.9121	0.8131	0.6986	0.8610	0.7469	0.3548	0.7799
	2	0.8412	0.7918	0.9077	0.8036	0.6987	0.8461	0.7540	0.3553	0.7705
	3	0.8476	0.8064	0.9096	0.8064	0.7025	0.8447	0.7581	0.3646	0.7715
C4	1	0.8577	0.8229	0.9096	0.8410	0.7930	0.8443	0.8041	0.4122	0.7734
	2	0.8869	0.8663	0.9385	0.8681	0.7951	0.9130	0.8162	0.4641	0.8147
	3	0.8655	0.8379	0.9208	0.8533	0.7837	0.8919	0.8054	0.4167	0.7722
R1	1	0.8492	0.7650	0.9175	0.8431	0.7542	0.9061	0.8409	0.4607	0.8293
	2	0.8471	0.7428	0.9185	0.8232	0.6934	0.8892	0.8381	0.4568	0.8180
	3	0.8459	0.7444	0.9142	0.8229	0.6992	0.8876	0.8434	0.4719	0.8256
	4	0.8846	0.7985	0.9530	0.8563	0.7473	0.9305	0.8736	0.5711	0.8773
R2	1	0.8621	0.7234	0.8954	0.8274	0.5782	0.8035	0.7837	0.2297	0.6513
	2	0.8967	0.7875	0.9375	0.8297	0.6066	0.8192	0.7808	0.2655	0.7020
	3	0.8707	0.7576	0.9203	0.8155	0.5724	0.7864	0.7941	0.2982	0.6935
R3	1	0.8929	0.8629	0.9131	0.8738	0.6162	0.7947	0.9348	0.1818	0.6986
	2	0.9114	0.8807	0.9299	0.8901	0.6269	0.7965	0.9380	0.2609	0.6442
	3	0.9134	0.8844	0.9371	0.8928	0.6724	0.7911	0.9370	0.2692	0.6623
R4	1	0.8828	0.8406	0.9148	0.8582	0.7657	0.8560	0.8659	0.3626	0.7067
	2	0.9284	0.9463	0.9622	0.8438	0.7219	0.8178	0.8687	0.3429	0.6771
	3	0.9014	0.8486	0.9326	0.8461	0.7167	0.8473	0.8696	0.3895	0.6839
Global Statistics	Min.	0.7844	0.7234	0.8767	0.7775	0.5724	0.7864	0.7469	0.1818	0.6442
	Max.	0.9284	0.9463	0.9622	0.8928	0.8033	0.9307	0.9380	0.7699	0.9319
	Mean	0.8602	0.8113	0.9187	0.8323	0.7196	0.8625	0.8255	0.4500	0.7801
	Median	0.8537	0.8019	0.9179	0.8277	0.7346	0.8636	0.8142	0.4145	0.7767

Both Models 1 and 2 only used PMS data, but Model 2 was fed with data from a shorter period. Nevertheless, Model 2 showed better results than Model 1, that presented better results for a couple of cases, namely for hotel R1 and for *Accuracy*, *Precision*, and *AUC* both for validation and test sets results, and for hotels R4 and C3 but only in a few of the metrics. For the remaining hotels, most metrics achieved higher results with Model 2. These differences show that more data does not always produce better models (Abbott 2014). As recognized by McGuire (2016), more data from the same source might not result in better models. This affirmation is particularly true if data does not have a significant causal relationship with the outcome, if data does not change significantly over time or if data lacks "quality".

Similarly, Model 3's results show that the introduction of additional features from other non-PMS data sources did not translate into better results for every hotel. For hotels C1, C2, and C4, Model 3 was beaten in every metric on both the validation and the test sets. On the contrary, almost all metrics for Model 3 and using the test set showed improved results over those of Model 2 for hotels R1, R2, R3, R4, and C3. However, this was not matched in the validation set, where the improvement did not homogeneously happen for all the metrics.

On the other hand, Model 4's results distinctly show that the inclusion of features specific to each hotel's characteristics and operations could impart substantial performance improvements. When compared with Model 3 test set results for R1, *Accuracy* increased over 3 percentage points, *Precision* over 10 percentage points, and *AUC* over 3 percentage points. For C1, both *Accuracy* and *AUC* increased over 3 percentage points while *Precision* increased over 2 percentage points.

From a general point of view, the overall statistics (Table 11) draw attention to some of the global results obtained. All metrics presented good results in terms of prediction performance (validation set). *Accuracy* ranged from 0.7775 to 0.8928. *Precision* ranged from 0.5724 to 0.8033. *AUC* ranged from 0.7864 (which is usually considered a fair to good model result) to 0.9307 (which is considered an excellent result). In terms of the generalization performance, i.e., the models' prediction capability on independent test sets (Hastie, Tibshirani, Friedman 2001), mean and median results show that for most hotels results were good. Nevertheless, this was not transversal to all of the hotels, particularly in terms of *Precision* and *AUC* for hotels R2, R3, and R4. These three were also the ones presenting the lowest cancellation ratio. This might indicate that, for low cancellation ratio hotels, either additional data or different features (as demonstrated by Model 4) should be added to try to improve the capture of cancellation patterns; or, it is just tough to predict cancellations for such hotels, maybe because cancelations have no pattern besides the costumer's own limitations.

Another important consideration arising from the results is the Pearson correlation values between Model 3's test set *Accuracy* values and the hotels' OTAs share, as well as between *Accuracy* and the hotels' cancellation ratio. The correlation between Model 3's *Accuracy* and the OTA's share in hotels can be considered moderate to strong (-0.5894). The correlation between Model 3's *Accuracy* and the hotels' cancelation ratio can be considered strong (-0.6282), suggesting the existence of a negative association between models' *Accuracy* and both the hotels' OTAs share and the hotels' cancellation ratio. When OTA's share or the cancellation ratio decrease, *Accuracy* increases, and vice-versa. Since there was also a moderate positive correlation between the OTA's share and the cancellation ratio, it is suggested that the higher the hotels' OTAs market share, the higher the cancellation ratio can be, and thus, the harder it is to predict cancellations accurately.

One of the powerful characteristics of XGBoost is the capability of generating three measures of each feature's contribution relative to the whole model: *Gain*, *Cover*, and *Frequency*. *Gain* measures the improvement in accuracy brought by a feature to the tree branches in it is on. *Cover* measures the relative number of observations for the feature. *Frequency* (also known as *Importance*) is a more straightforward measure that is calculated by counting the number of times a feature is used in all generated trees. This count means that a feature with a *Frequency* of 0 (zero) was not used in the model. The *Frequency*/*Importance* in Model 3 shows which features were used in each hotel's model version (Table 12). As usual with predictive modeling, not all features had substantial influence in the prediction of the outcome (Hastie, Tibshirani, Friedman 2001). From the 29 features, only 13 to 15 features were used, depending on the hotel. Unexpectedly, only PMS originated features. Features from the other data sources were not used. As previously described, the inclusion of features from non-PMS data sources induced minimal performance improvements for some of the hotels. Further, the improvements were not due to the information gain brought to models by those features but due to the way the XGBoost algorithm works. The parameters used to control overfitting allowed the tuning of the model's complexity, making it simpler and less likely to overfit. Consequently, the introduction of features from other data sources, although not adding more information, made some models more robust to noise.

Table 12 - Features employed per hotel model (Model 3)

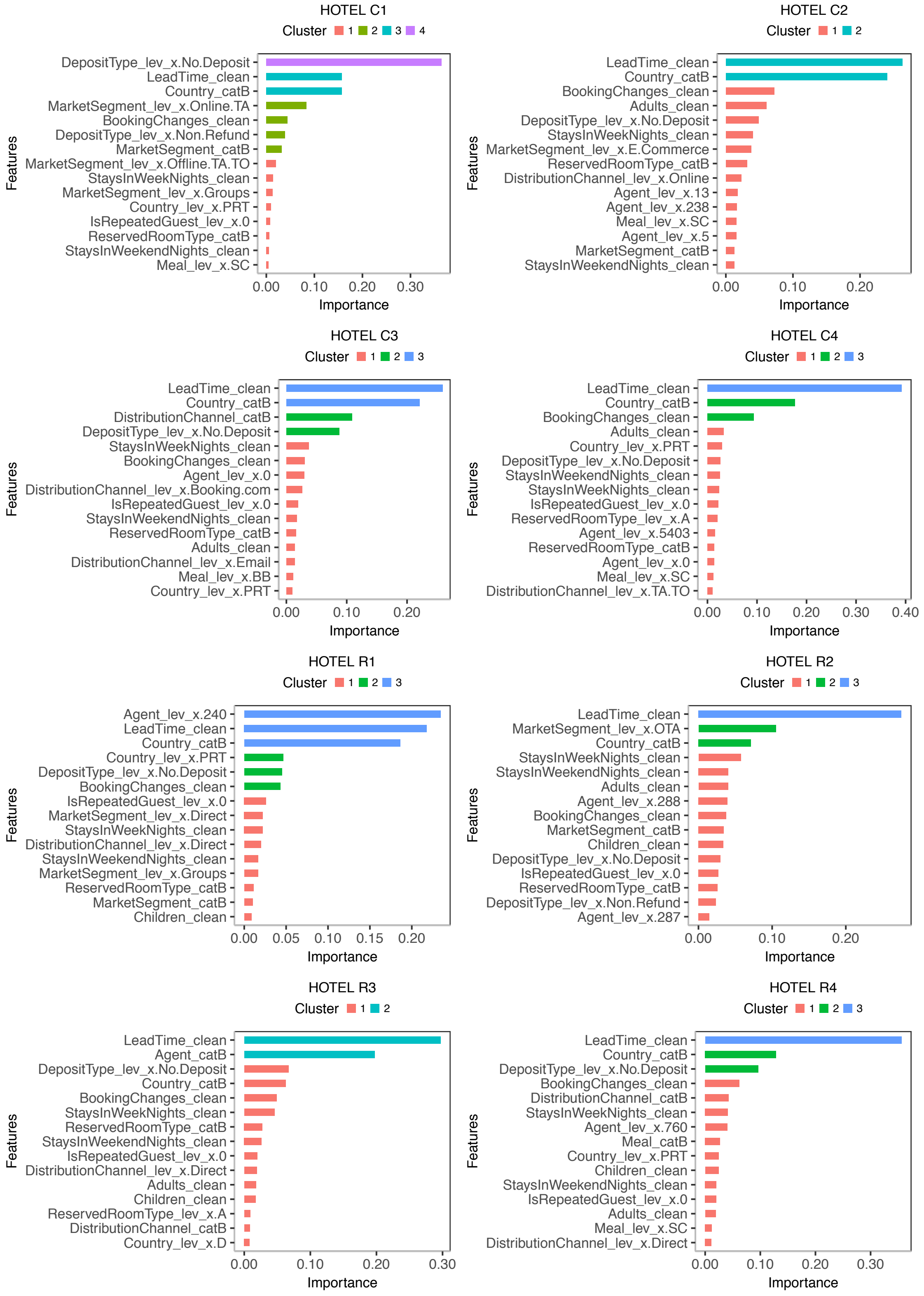
Feature	**C1**	**C2**	**C3**	**C4**	**R1**	**R2**	**R3**	**R4**
Adults	•	•	•	•	•	•	•	•
Agent	•	•	•	•	•	•	•	•
AvgQuantityOfPrecipitationInMM								
Babies	•	•		•	•	•	•	•
BookingChanges	•	•	•	•	•	•	•	•
Children	•	•		•	•	•	•	•
Company							•	
Country	•	•	•	•	•	•	•	•
CustomerType								
DayOfYear								

Feature	C1	C2	C3	C4	R1	R2	R3	R4
DaysInWaitingList								
DepositType	•	•	•	•	•	•	•	•
DistributionChannel	•	•	•	•	•		•	•
HotelsWithRoomsAvailable								
IsRepeatedGuest	•	•	•	•	•	•	•	•
LeadTime	•	•	•	•	•	•	•	•
MarketSegment	•	•	•	•	•	•		•
Meal	•	•	•	•	•	•	•	•
nHolidays								
PreviousCancellationRatio								
RatioADRbyCompsetMedianDifference								
RatioMajorEventsNights								
RatioMinorEventsNights								
ReservedRoomType	•	•	•	•	•	•	•	•
StaysInWeekendNights	•	•	•	•	•	•	•	•
StaysInWeekNights	•	•	•	•	•	•	•	•
ThirdQuantileDeviationADR								
TotalOfSpecialRequests								
WorseThan								
29 Features (without features for specific categorical levels)	**15**	**15**	**13**	**15**	**15**	**14**	**15**	**15**

The analysis of the top 15 most important features per hotel, based on the *Frequency*/*Importance* measure calculated by XGBoost, is depicted in Figure 46 and shows that the order of importance differed substantially by the hotel. These differences were not only in the ranking order of each feature, but also on the features that composed the top 15. Because XGBoost employs one-dimensional clustering to determine the grouping of features in terms of importance, it is possible to verify that there are differences between hotels in terms of the number of clusters and the number of features in each of the clusters, as well as in the degree of importance of the feature by cluster and by hotel. However, some features had similar importance for every hotel. *LeadTime* was the most important feature for six of the hotels and the second most important in the other two. From these and for hotel R1, a feature that represents bookings from a specific level (240) of the *Agent* categorical feature had a higher importance. In C1, the most important feature represented the level “No deposit” of the categorical feature *DepositType* jointly with the level “Non-refundable”. *Country* was also one of the most important features for every hotel, except for R3 where it came in fourth. For all other hotels, *Country* usually came in second or third place.

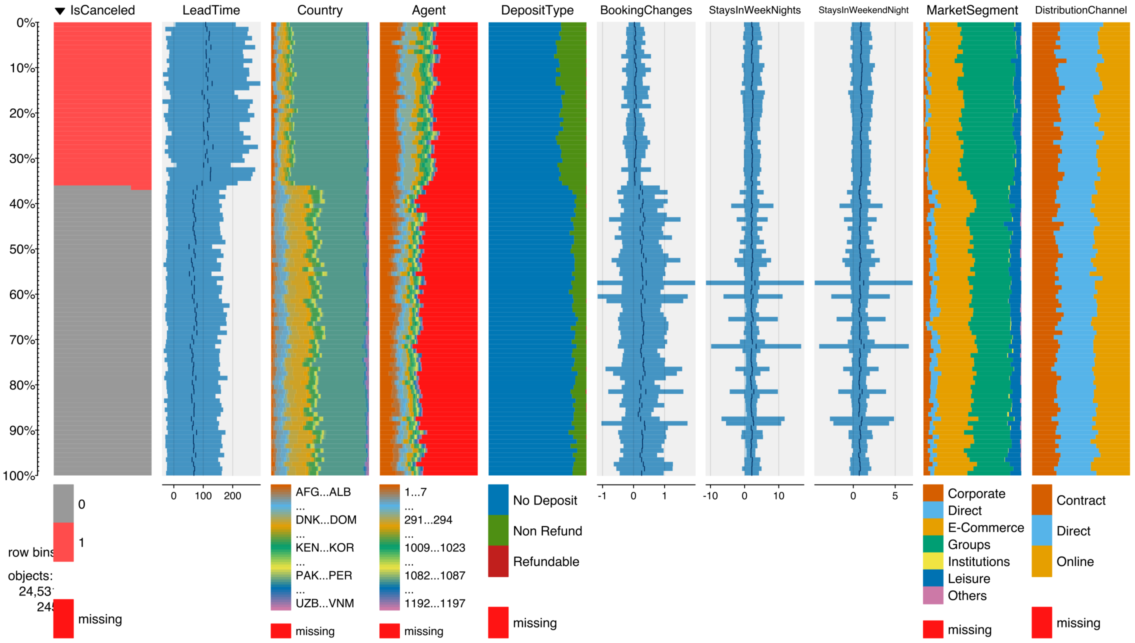
Another feature of high importance for all hotels was *BookingChanges*. Interestingly, Figure 46 highlights that the feature *StaysInWeekNights* was more important in the cancellation prediction than the feature *StaysInWeekendNights*, except for in C4 where the results were not distinguishable.

Figure 46 - Top 15 features per hotel (Model 3)



Identifying which features are more important in the prediction of the outcome of a booking allows to narrow down cancellation drivers. A smaller number of dimensions can make it easier to study the data and uncover hidden patterns. For example, Figure 46 presents a “Tableplot”, a powerful big data visualization technique that allows the exploration and analysis of large multivariate datasets (Tennekes, de Jonge 2017). The most important predictive features for hotel C2 and Model 3’s dataset is represented in this plot. At a glance, it is possible to verify that average *LeadTime* tended to be higher for canceled bookings. Other patterns that also stood out in canceled bookings were: (1) lowest average number of amendments to bookings (*BookingChanges*); (2) higher average number of adults per booking; (3) higher percentage of “Non-refundable” bookings (*DepositType*); (4) higher number of stays over weekends (*StaysAtWeekendNights*); (5) higher number of “Groups” and lowest “Leisure” customers (*MarketSegment*); and (6) bookings for room type “A” canceled more than bookings for other room types (*ReservedRoomType*). Although these patterns require more in-depth analysis, this is a starting point to understand the reasons behind cancellations and to define measures to prevent them, or at least, to better estimate them.

Figure 47 - Top predictive features visualization (Hotel C2 - Model 3)



Comprehending which features are the best descriptors for possible cancellations allows hoteliers to rethink their cancellation policies in different ways. Restrictive cancellation policies reduce demand while less restrictive policies, in addition to boosting demand, improve revenue due to the application of lesser discounts. Why not take advantage of the fact that a large part of hotel distribution is now made online and encourage the application of dynamic cancellation policies? Why not foster the application of cancellation policies that vary according to the lead time, country of origin, or staying days of the week?

The identification and comprehension of the importance of features regarding booking cancellations require hotels to have quality data to better support decisions. Without quality data, models like the ones presented here could not be built, or at least not with as good results. Sometimes, the lack of quality is the result of the human side of data. How and when data entered at the different systems, like for the classification of a booking market segment, is often carried out by a human operator. If the hotel/brand does not have clear rules on how to classify bookings, this is done at the operator's discretion, resulting in a worthless classification. Another example is the time between the delivery of the booking to the hotel and the time the booking is entered in the system. Although many bookings are automatically inserted into the hotel's PMS via different electronic interfaces (e.g. "Channel manager" or CRS), depending on the hotel/chain, some bookings are still entered manually. If operators do not enter bookings in the PMS at the day of their delivery to the hotel or do not enter the correct delivery date when the booking is created, one of the most important features in terms of cancellation prediction, *LeadTime*, will be negatively influenced (in terms of quality speaking). This manual entering of bookings in the PMS system is what happens in hotels R2 and R3: at the time of the period of study, these hotels did not use a "Channel manager" to integrate their electronic bookings automatically in the PMS. Therefore, when in times of high workload, hotel operators tend to enter in the PMS bookings for near arrival dates and postpone the entering of bookings for more distant dates. It is possible that this may explain part of the not so good performance of the models for these hotels.

4.2.6 Deployment

During the construction of the final models, developments on PhD funding enabled the implementation of field tests to evaluate the deployment concepts developed during the construction of the exploratory models. Consequently, deployment of final models was the subject of the field tests that are the subject of the following chapter.

4.3 Discussion

In this chapter, it was possible to combine PMS data with data from other sources to extend the work described in Chapter 3 to answer both RQ1 and RQ2. The use of data from eight hotels, four of them with a different type classification from than the ones used in the exploratory models, not only confirmed the indications given by the exploratory models about RQ1, it also sanctioned the generalization of the results. However, despite the enormous potential of big data for the hotel industry, as shown by the results, the inclusion of data from multiple sources did not produce significant performance improvements.

Features explanatory power not always imply predictive power (Domingos 2012; Shmueli, Koppius 2011). Descriptive statistics have shown this, at least for some of the hotels and for some of the features from non-PMS data sources. Features like the existence of events on the period of stay, or if the period of stay covered holidays, and social reputation or price positioning compared to competitors, although significantly associated to bookings cancellation, this association was not reflected in the predictive power of features. As already proposed by Pan, Yang (2017b), this raises the question of whether the use of big data is justifiable in hospitality research. A low-performance impact does not always justify the costs associated with collecting, storing, and processing data, as well as the time required to process large volumes of data or the time spent in data preparation and modeling. Therefore, the application of big data requires a thoughtful study of the associated costs and benefits. As shown, models that used features based only on the hotel's PMS data performed better than those that included features from multiple sources. In fact, features from data sources other than the hotel's PMS were never included in the ensemble of decision trees in any of the models, for any of the hotels.

The identification and comprehension of feature importance in terms of booking cancellations strengthen the need that hotels should strive to have quality data. Without quality data, models like the ones presented here are not possible to build, or at least will not produce good results.

Even though the results of the best models, Models 3 and 4, did not surpass the results obtained by the exploratory models, they were more robust. This robustness can be perceived by the results of the test set which, unlike the exploratory models and similar works (Huang, Chang, Ho 2013; van Leeuwen 2018), did not intersect the training set. Additionally, because PMS' datasets did not contain bookings' values at extraction time but values before check-in/cancellation time, training data were suitable for the model to predict the outcome of bookings from an unknown period. A sign that the time at which features are extracted has an impact on model's performance is the fact that *BookingsChanges* is one of the features with more predictive importance in all hotels. Thanks to these contributions, when compared with exploratory models, the new models were less likely to capture noise in data and could generalize better.

Overall, final models' performance results reinforce the confirmation given by the exploratory models that hotel's PMSs are a good data source to extract and derive features for machine learning models to predict booking cancellations with high accuracy. Concurrently, Model 4 results emphasize that the best models are attained with the inclusion of features that capture each hotel's characteristics and operating environment. Undeniably, this answers RQ1 positively. In other words, it confirms that detailed booking data based on the hotel's PMS are more relevant to predict booking cancellations than data in the PNR format. On the other hand, the study of features' predictive importance demonstrated that the inclusion of features from other sources did not improve models' performance, thus answering RQ2 negatively.

Based on the preliminary results coming out of the elaboration of the final models, field tests were designed and put into production to evaluate the performance of these models in a real work environment. Tests in a real work environment enabled the assessment of the models' accuracy, understanding of existent deployment problems, and understanding of operational benefits. The design of the field tests, its challenges, performance results, and business impact will be described in the next chapter.

4.4 Summary

Like for the development of the exploratory models, CRISP-DM proved to be an outstanding method for building predictive models. The iterations between the different phases of CRISP-DM allowed a good comprehension of both the business and the data, promoting the development of better modeling datasets. The quality of these datasets in turn, allowed the development of predictive models that achieve good results.

The use of detailed booking data from two different type of hotels, four resort hotels' PMS, and four city hotels, from the period of January 1st, 2016 to November 20th, 2017, with cancellation rates ranging from 12.2% to 40.0%, proved to be a good choice to understand the similarities and dissimilarities between hotels and to build the predictive models.

The application of data science tools and capabilities, such as data mining, data visualization, descriptive statistics, and statistical tests to PMS data combined with data from other sources, namely for events, holidays, online prices and inventory, social reputation and weather forecast, made it possible to demonstrate which features are the most appropriate to predict booking cancellations. Data science tools and capabilities also helped demonstrate how each hotel's operation characteristics (for example, each hotel dependency on OTAs as a distributor) has an influence on cancellations.

XGBoost confirmed to be an excellent classification algorithm in three fronts: 1) in terms of the quality of results achieved; 2) in terms of processing speed; 3) in terms of understanding each feature predictive importance.

Using R server on HDInsight platform enabled multi-processing and distributing computing to accelerate the computing tasks, which, in this particular case, were very demanding operations, especially in Models 3 and 4, as it involved the merging of PMS data with data from the other sources, some of them requiring the processing of millions of observations.

The good results achieved confirmed that features constructed from hotels' PMS data sources had much more potential than PNR format features, thus answering affirmatively to RQ1. Conversely, the study of the features' predictive importance showed that features from additional data sources did not contribute directly to model's improvement, thus answering negatively to RQ2.

5 FIELD TESTS

Notwithstanding final models' results not showing big data could directly contribute to the improvement of the performance of bookings cancellation prediction models, it showed how changes in data collection, feature engineering, modeling methods and in algorithms could improve models' performance. The question now was whether the good performance results could be maintained in a production environment.

Since the beginning, the project included the requirement to test the models in a real production environment and with that assess their impact on business (RQ3). However, this requirement was also one of the major risks of the project, because it required two essential prerequisites to be fulfilled. First, that hotels agreed to participate and commit resources to the tests, and second, the existence of computing resources that made possible the operationalization of the tests. The second of the prerequisites was achieved in January 2017, with a grant endowed by Microsoft of the Azure Data Science Award. The grant had to be spent in the usage of Azure resources during 2017 and was not extensible in time or value. Therefore, based on the results from the exploratory models and the final models' preliminary results, it was decided to contact the same hotels to understand if any was willing to participate in field tests. With the agreement from two of the hotels, the development of a prototype started in February 2017, in parallel with the collection of data from non-PMS data sources, for the development of the final models.

The application of machine learning to build predictive models, in the context of quantitative empirical modeling, i.e., *"building and assessment of a model aimed at making empirical predications"* is known as "predictive analytics" (Shmueli, Koppius 2011, p. 555). The description of what is and why predictive analytics can help answer RQ3 is the subject of the introductory section of this chapter. The following section presents a comprehensive description of how the prototype was built and the materials used. This section is followed by a presentation of the results and its discussion. Lastly, a global discussion on the impact of this prototype, followed by a summary of the work done.

5.1 Introduction

Predictive analytics models comprise two components: empirical predictive models developed to predict new/future observations and methods for evaluating the predictive power of these models. Predictive analytics has a vital role in theory building, theory testing, and relevance assessment. Scientific research in predictive analytics can assume different roles: generating new theoretical results, development of evaluation measurements, comparing competing theories, improving existing models, assessing relevance or assessing predictability (Shmueli, Koppius 2011). Shmueli and Koppius (2011) conducted a literature survey to investigate to what extent was predictive analytics integrated into empirical Information Systems research and concluded that only seven of 52 papers with predictive claims employed predictive analytics. This shortage of studies on the subject is also recognized by other authors, who state that the development of successful predictive analytics applications is not addressed in textbooks, not even in the form of general principles how should be deployed (Abbott 2014; Domingos 2012). The gap in the application of predictive analytics may be explained by the difficulty to overcome obstacles for its operationalization (Abbott 2014): management (a shift in resources allocation and mentalities is required), data (existence of quality data on the subject), modeling (model complexity issues), deployment (integration and practicality issues). Although some authors recognize the importance of predictive analytics in the hospitality industry, especially for the discipline of revenue management (Cross 2016; Yeoman 2016), studies that address predictive analytics applications specific to hospitality are lacking.

As described in Chapter 2, despite the importance of predicting hotel bookings cancellation, not many studies have addressed the subject. It is not surprising, therefore, that no study until now addresses this problem from an empirical perspective, that is, from a predictive analytics research perspective. To answer this gap, and simultaneously answer RQ3, this chapter shows how an RMS component prototype could be built and implemented. The prototype consists of a machine learning model that uses PMS data to predict which hotel bookings have a high likelihood of being canceled. The prototype was deployed in two hotels in order to assess its performance in a real production environment. The deployment incorporated active hotel actions to prevent cancellations of bookings predicted to cancel with high probability, which has also been the subject of evaluation.

Details on the construction of the prototype, its deployment, application in daily operations, and results assessment are presented in the following section.

5.2 Materials and methods

Economic theories such as rationing, free entry, price discrimination, and monopoly pricing provide insights that are essential to revenue management. Certain economic fundamentals and assumptions serve as the basis of revenue management in the hospitality industry, namely, product perishability, limited capacity, high fixed and low variable costs, unequal demand over time, possibility to forecast demand, possibility to segment demand, and different price elasticities of market segments (Talluri, Van Ryzin 2005; Ivanov 2014). Nevertheless, revenue management practice often diverges from classical economic theory in important aspects (Talluri, Van Ryzin 2005). For example, the application of price elasticity demand theories in the hospitality industry is more theoretical than practical. For instance, customers can always change to a different hotel if the price increases or due to brand loyalty stay at a hotel even when the price in other hotel decreases (Ivanov 2014). This gap between theory and application renders the empirical evaluation of a machine learning model to predict hotel bookings' cancellations an undeniable challenge that should be addressed in the context of Design Science Research (DSR). DSR requires the development of an artifact, in this case, a prototype of a RMS component, which fulfils the two requirements of DSR: relevance—by addressing a real business need—and rigor—by applying the proper body of knowledge in the artifact development (Cleven, Gubler, Hüner 2009; Hevner, March, Park, Ram 2004). In this case, this body of knowledge is encompassed by data science fields: computer science (machine learning, databases, and data visualization), statistics and domain knowledge (O'Neil, Schutt 2013; Flath, Stein 2018).

CRISP-DM was again the process model employed for the development and assessment of the models. However, due to the specific nature of these models including deployment and fully automated data collection and preparation, the structure of the sub-sections of this section does not follow the same pattern of the previous two chapters.

5.2.1 System design

The system has several different objectives: the automatization of the modeling tasks; to deliver information for the hotel to act upon; and to register information that enables it to assess the performance of the booking cancellation prediction model in a real production environment. The system was designed based on the following requirements and specifications:

- For modeling:
  - The system trains daily with a dataset of all reservations on-the-books, enabling it to learn with changes in bookings and changes of patterns that occur over time.
  - Each day, the system builds a new model and automatically executes hyper-tuning of parameters, whose performance is compared with the performance results of the previous seven days. This evaluation supports a decision for replacing the current model parameters to be replaced with the new ones or continue to use the previous parameters.

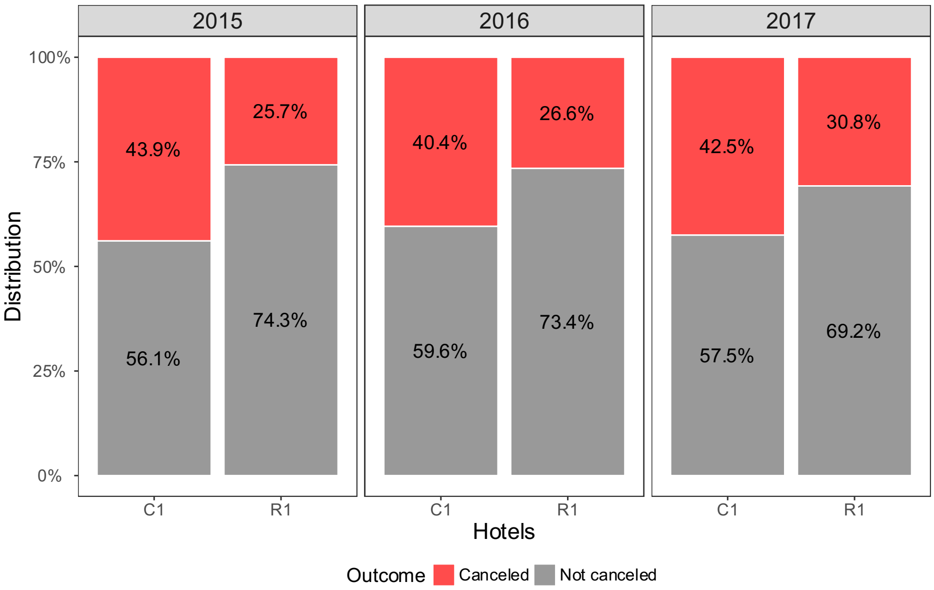
  - The predictions and performance results of the preceding days are stored in a database for evaluation, and where applicable, reused as model elaboration features.
  - The system trains by incorporating the incorrect predictions of previous days as penalizations and the correct predictions of previous days as rewards, with costs being class-dependent (false positives have higher costs than those of other miss-classifications).
  - 50% of the new bookings should be marked as the "control group", indicating that the details of these bookings would never be shown to hotels thus enabling A/B tests.
  - Global demand and net demand for future dates are calculated based on existing bookings and model prediction results.
- Usability:
  - A web-based platform with a visualization component should be accessible by hotel staff and researchers anywhere at any time.
  - Hotels should have a login per staff user to access the application.
  - Every action executed by hotel staff should be logged.
  - Global totals, totals per room type of demand, and net demand are displayed in a planning screen.
  - Details of bookings that were identified as likely to cancel (and not part of the "control group") for the current date or previous days should be available for consultation.
  - Booking attributes that may lead to the identification of customers should be not be displayed or recorded by the system (to enable research purpose usage).
  - The system should report the actions made toward bookings that were identified as likely to cancel to prevent their cancellation.
  - The system must provide the visualization of the model performance results daily.
  - The system must provide the analysis of model predictions and effective performance results without disclosing the results of the A/B testing.

5.2.2 Hotel participation, data understanding, and data description

Convincing hotels to participate in the assessment was challenging as hotels were required to commit resources to the project, particularly human resources. Hotels' staff were required to use the prototype on a daily basis and incorporate the prototype predictions in their demand-management decisions. Hotels' staff were also required to analyze the bookings that were predicted as likely to cancel and decide which customers to contact to try to prevent a cancellation.

Two of the previously studied hotels, belonging to the same hotel chain, C1, and R1, accepted to participate. Both hotels have more than 200 rooms and are classified as four-star hotels. Data were available from July 2015 to August 2017. Because C1 was engaged in a soft-opening process until the end of August 2015, only data from September 2015 onwards was considered for modeling C1. Figure 48 presents the cancellation ratios of both hotels, which oscillate between 25.7% in 2015 and 30.8% for 2017 for R1 (until August 2017) and exceed 40% for C1. Values are slightly superior to the values of these hotels that were shown for the final models' datasets. This difference happens because, for the new prototype datasets, bookings with a cancellation outcome date outside the period of study for arrivals (July 2015 to August 2017) were not removed from the dataset. Unlike the final models, in this case, only PMS data was used. This decision was made considering two factors. First, final models' preliminary results showed that PMS was the primary source of features with predictive power. Second, because time limitations did not allow the collection of enough data from the non-PMS sources.

Figure 48 - Cancellation ratio per year



C1 and R1 PMS datasets structure and level of detail were the same as the ones in the datasets studied in the final model's development.

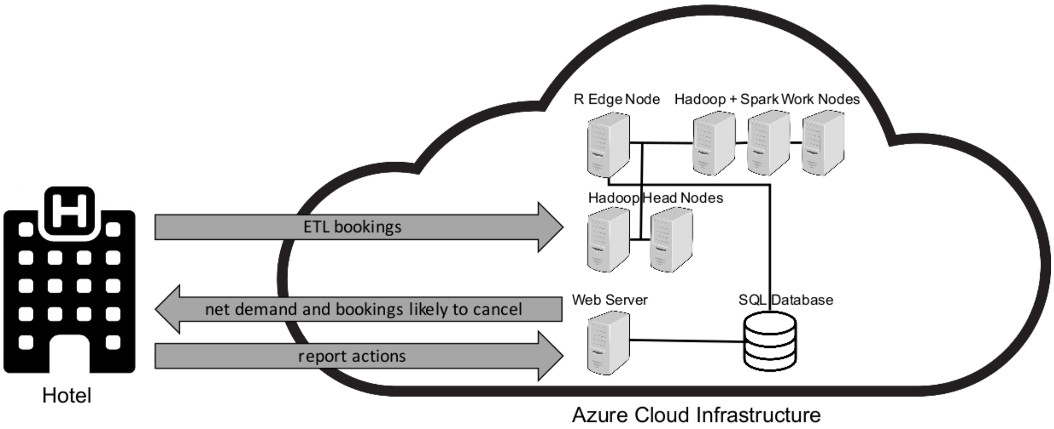
Regarding data preparation, having in account what was known with the development of the final models, the features included in the modeling datasets were almost the same as the ones included in final models 1 and 2. However, the going back and forth in the processing phases showed that there was some leakage. To solve this problem the *Country*, *RequiredCarParkingSpaces*, and *ReserverdRoomType* were removed. Data exploration revealed that in the last months the hotels started to work more with "waiting lists" for peak demand dates. Therefore, the feature *WasInWaitingList* was reintroduced in these modeling datasets. One feature that was not included in the modeling dataset and that it proved to be useful in the final models 1 and 2 was *DayOfTheYear*. It was not included because its contribution was only found later on, in the development of those final models. Lastly, the feature *LeadTime* was replaced by an engineered feature called *LiveTime*. This new feature (Appendix C) was created to capture the time elapsed between *LeadTime* and the booking outcome date, or in other words, the number of days that the booking was “alive”. Thus, for non-canceled bookings (“A:” type bookings”) it holds the *LeadTime* value. For canceled bookings (“B:” type bookings) it holds the number of days between creation date and cancellation date. For bookings due to arrive (“C:” type bookings) it holds the number of days between its creation date and the current date (processing date).

5.2.3 System architecture and modeling

To comply with all previously mentioned prototype requirements and specifications, and to render the system technically reliable and capable of adequate performance, the system was built on top of the Microsoft Azure cloud platform for taking advantage of several open-source components and technologies available as services in Microsoft Azure (see Figure 49):

- One HDInsight Linux based, Hadoop and Spark cluster with R Server. This component enabled Hadoop/Spark-based big data processing, R to be used in the Spark context and took advantage of XGBoost performance efficacy by using the cluster capabilities to distribute the processing among the different machines.
- One SQL database to process and store logs for all operations. This component also stored all prediction results with actions of the users. The database structure and database summary statistics are described in Appendix F.
- One web server. This component published the visualization layer in the form of a dynamic website, built in C# and asp.net. In this website, users can consult demand, predictions, and report the actions made for bookings identified as likely to cancel.

Figure 49 - System architecture diagram

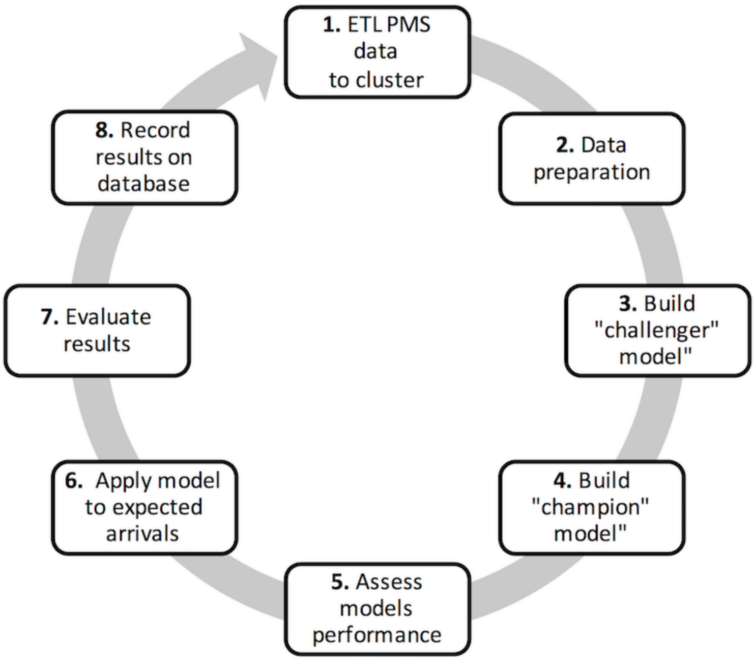

R Edge Node
Hadoop + Spark Work Nodes
Hadoop Head Nodes
ETL bookings
net demand and bookings likely to cancel
report actions
Web Server
SQL Database
Hotel
Azure Cloud Infrastructure

Since each hotel had each own PMS database located in servers at the hotels’ premises, a fully automated Extract, Transform and Load (ETL) process was created in each of the hotels for daily extracting all bookings from the hotels PMS’, transforming the data into a CSV dataset file and loading the data into the Hadoop cluster.

Models do not stay effective indefinitely. Their performance tends to worsen with time. The main reason models’ performance deterioration is “concept drift” (Abbott 2014; Gama, Medas, Castillo, Rodrigues 2004; Webb, Hyde, Cao, Nguyen, Petitjean 2016). To overcome this vulnerability and to enable the system to learn from new data continuously, the system was designed to incorporate the “Champion-challenger” approach (Abbott 2014). Rather than waiting for a decrease in model performance to build a new model, a challenger model is built daily, and its performance is compared with the performance of the current model. The model with superior results will be selected. This fully automated daily cycle, which is illustrated in the diagram of Figure 50, is composed of eight steps:

1. **ETL PMS data to cluster**: at a predefined time, SQL jobs extracts all bookings from the PMS database, transforms data to the format required by the modeling component and loads the data to the Hadoop cluster via a Windows Powershell script.
2. **Data preparation**: this important step includes the selection of data, definition of the training and testing datasets, removal of the unused features, data cleaning, construction of engineered features, reformatting of categorical features, and calculation of a weight per booking/observations (as next explained). Due to the dataset shift/concept drift problem, once again convenience splitting was used to split data into monthly time blocks, but this time only in training and testing sets. The reason for not creating a validation set had to do with the fact that the validation set would be composed with the “C:” type bookings.
3. **Build “challenger model”**: using the training dataset, a ten-fold cross-validation mixed grid/random-search is executed to hyper-tune model parameters (using the code presented in Table 10). The model is trained with the selected hyper-tuned parameters.
4. **Build “champion model”**: train a model with the parameters employed on the previous day.
5. **Assess models’ performance**: in this step, both models are fed with the testing set and both *Accuracy*, and *AUC* metrics values are compared. When the “challenger” model outperforms the “champion” model for the last seven days’ average and on at least four of the days for both metrics, the “challenger” is selected to be the model to use. Otherwise, the “champion” model will continue to be used.
6. **Apply the selected model to expected arrivals**: this step involves the application of the selected model to all future arrivals (“C:” type bookings) and predicts their outcome.
7. **Evaluate results**: (both models) calculation of classic machine learning performance metrics (*Accuracy*, *AUC*, *Precision*, *F1Score*, *Sensitivity* and *Specificity*), regarding both the training datasets and the testing datasets. Calculate the ratio of predicted bookings as likely to cancel for future arrivals (“C:” type bookings).
8. **Record results in database**: all performance metrics and all predictions of the current day are recorded in the database to enable further analysis and enable the use of previous predictions in the creation of the weighting mechanism.

Figure 50 - Daily automation cycle program


1. ETL PMS data to cluster
2. Data preparation
3. Build "challenger" model"
4. Build "champion" model"
5. Assess models performance
6. Apply model to expected arrivals
7. Evaluate results
8. Record results on database

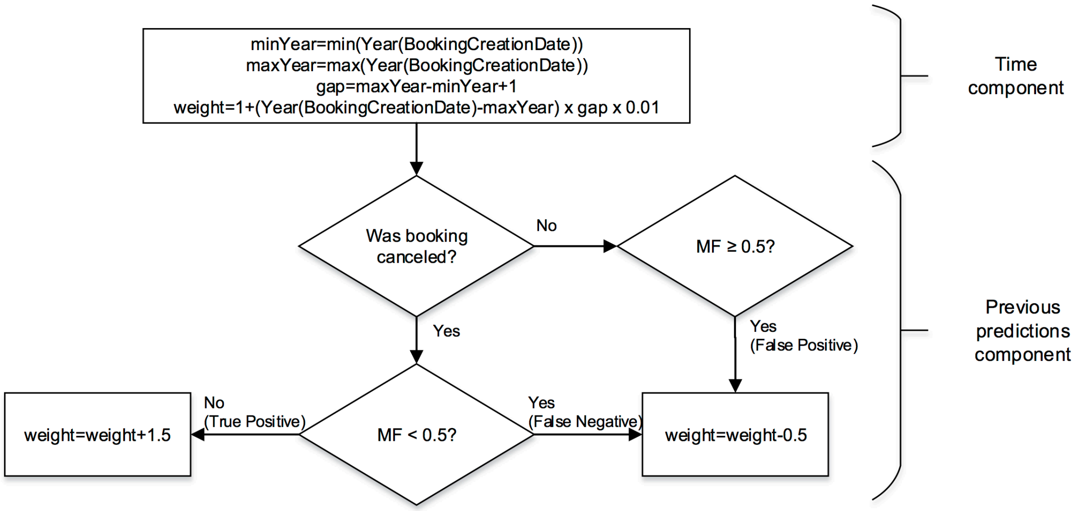
Note that, since cancellation patterns change over time and because the system was required to learn continuously, a weighting mechanism was created to attribute higher importance to recent bookings and to incorporate cost-sensitive learning by example based on previous predictions hits and errors (Abe, Zadrozny, Langford 2004). In fact, hotel bookings are dynamic, i.e., over time there is a change in bookings' attributes (e.g., arrival date, length of stay, number of persons, among others). On the other hand, time to arrival influences cancellations: a booking can be predicted as "likely to cancel" in one of the days, but as "not likely to cancel" on the next day. Measuring the precision of previous predictions on unstable observations required the development of a new measure, Minimum Frequency (MF):

$$MF = \frac{\sum_{i=1}^{n} \hat{y_i}}{n} \qquad (2)$$

In the MF formula, *n* is the number of days since the booking has arrived at the hotel and has been processed by the predictive system and $\hat{y_i}$ is the prediction classification for each day *i* it was processed. The prediction is binary: 0 for classified as "not likely to cancel" or 1 when classified as "likely to cancel".

As illustrated in Figure 51, the weighting mechanism is comprised of two components. The "time component" calculates the base weight according to the booking antiquity. Then, the "previous predictions component" uses the booking outcome status and the MF measure to assign a penalization to every false negative and false positive observations on the dataset, or a bonus to true positive predictions. The MF threshold to classify if the prediction was correct was set to 0.5.

Figure 51 - Observations weighting mechanism diagram



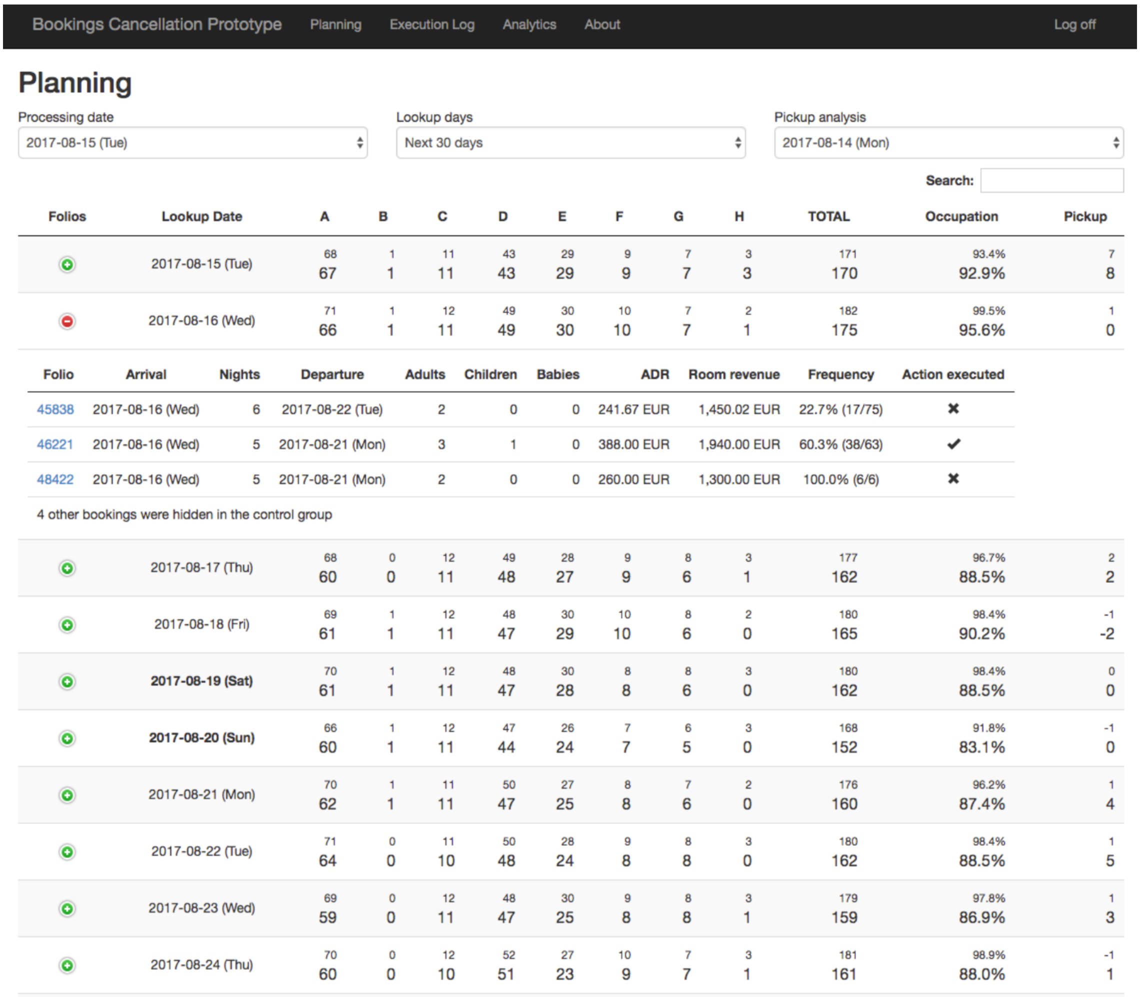
5.2.4 Development and deployment

The main component of this system prototype—the modeling component—was written in R and continuously run in the R Edge node of the HDInsight cluster. Every day, at a predefined hour, this component executed the daily automation cycle described in Figure 50. This modeling component and its visualization component were deployed in April 2017. After a set of tests, adaptations, and optimizations, the system was made available to hoteliers on May 1st, 2017. However, it was not until the end of May that hotels started to use the prototype systematically. Initially, the evaluation period was defined to run from June to September 2017. However, due to hotel human resources constraints, this period had to be shortened to August 2017.

An initial kickoff meeting was held in April to provide training to hotel users (revenue management team) about the visualization component of the system. The training explained how users should report actions to prevent cancellation of bookings signalized, consult logs and analyze modeling performance results. The training also discussed how to visualize a planning for future dates and how to identify bookings that were predicted as likely to cancel. The main screen of the prototype visualization component (planning for future dates screen, Figure 52) enables users to visualize the demand for each room type (smaller font) and the net demand (larger font) for current and future dates one year in advance. The net demand is calculated by deducing the total number of bookings that were predicted to be canceled. The planning also exhibited the daily totals of demand, occupation ratios, and pickup (difference in the total bookings between a date - the previous day by default - and the day of the visualization). A button on each one of the day lines enables users to check the PMS identification (*Folio number*) of the bookings that were identified as likely to cancel. The button also allows the visualization of additional information, including booking attributes such as arrival date, nights, departure date, number of persons, ADR, total room revenue and frequency. Frequency was a metric that was necessary to create to show users the number of days in which the booking was identified as likely to cancel in relation to the total number of days that the booking was processed by the system (Figure 52). Since A/B testing was used for system assessment, 50% of the bookings were used as a control group ("A" group), and the remaining 50% of the bookings were used as the verification group ("B" group). Therefore, users could only view the details of bookings predicted as likely to cancel from the "B" group. A click on the *Folio number* enables users to report to the researchers the actions that were taken to avoid a booking cancellation, including how the action was executed and what was offered to (or asked of) the customer.

Prototype pages for reporting customer's contacts, for evaluating the daily predictions and analyzing models performance are shown in Appendix G.

Figure 52 - Prototype's main screen - Planning



Planning

Processing date: 2017-08-15 (Tue)

Lookup days: Next 30 days

Pickup analysis: 2017-08-14 (Mon)

Search:

Folios	Lookup Date	A	B	C	D	E	F	G	H	TOTAL	Occupation	Pickup
⊕	2017-08-15 (Tue)	68 67	1 1	11 11	43 43	29 29	9 9	7 7	3 3	171 170	93.4% 92.9%	7 8
⊖	2017-08-16 (Wed)	71 66	1 1	12 11	49 49	30 30	10 10	7 7	2 1	182 175	99.5% 95.6%	1 0

Folio	Arrival	Nights	Departure	Adults	Children	Babies	ADR	Room revenue	Frequency	Action executed
45838	2017-08-16 (Wed)	6	2017-08-22 (Tue)	2	0	0	241.67 EUR	1,450.02 EUR	22.7% (17/75)	✖
46221	2017-08-16 (Wed)	5	2017-08-21 (Mon)	3	1	0	388.00 EUR	1,940.00 EUR	60.3% (38/63)	✔
48422	2017-08-16 (Wed)	5	2017-08-21 (Mon)	2	0	0	260.00 EUR	1,300.00 EUR	100.0% (6/6)	✖

4 other bookings were hidden in the control group

Folios	Lookup Date	A	B	C	D	E	F	G	H	TOTAL	Occupation	Pickup
⊕	2017-08-17 (Thu)	68 60	0 0	12 11	49 48	28 27	9 9	8 6	3 1	177 162	96.7% 88.5%	2 2
⊕	2017-08-18 (Fri)	69 61	1 1	12 11	48 47	30 29	10 10	8 6	2 0	180 165	98.4% 90.2%	-1 -2
⊕	**2017-08-19 (Sat)**	70 61	1 1	12 11	48 47	30 28	8 8	8 6	3 0	180 162	98.4% 88.5%	0 0
⊕	**2017-08-20 (Sun)**	66 60	1 1	12 11	47 44	26 24	7 7	6 5	3 0	168 152	91.8% 83.1%	-1 0
⊕	2017-08-21 (Mon)	70 62	1 1	11 11	50 47	27 25	8 8	7 6	2 0	176 160	96.2% 87.4%	1 4
⊕	2017-08-22 (Tue)	71 64	0 0	11 10	50 48	28 24	9 8	8 8	3 0	180 162	98.4% 88.5%	1 5
⊕	2017-08-23 (Wed)	69 59	0 0	12 11	48 47	30 25	9 8	8 8	3 1	179 159	97.8% 86.9%	1 3
⊕	2017-08-24 (Thu)	70 60	0 0	12 10	52 51	27 23	10 9	7 7	3 1	181 161	98.9% 88.0%	-1 1

To prevent cancelation of bookings that were identified as likely to cancel, the hotel revenue management team had *carte blanche* from the hotel chain board to offer any services or discounts they deemed suitable according to the booking potential revenue loss. These discounts included breakfast discounts (to customers who have booked room-only rates), free room-type upgrades or discounts on room-type upgrades, free meals or discounts on meal packages, and discounts on other services such as car parking, SPA treatments and free tickets for local attractions.

Initial contacts with customers revealed this type of approach (offering discounts or complementary services) to be very demanding in terms of human resources cost besides being financially costly. Contacted customers decided to request additional discounts (e.g., when offered a 20% discount on breakfast, customers would ask for free car parking), which could result in higher costs/less margin and above all, was a highly time-consuming task. Therefore, the hotel revenue management team rapidly decided to change the contact policy and, with the researchers agreement, decided to inquire customers on details that might improve the quality of service, such as: the type of bed preferred, the expected hour of arrival to ensure that rooms could be prepared in a timely manner, children's ages (for the size of beds/cots), car license plate (to accelerate the check-in process) or credit card details, when the customers had not fill the credit card details or the data was not validated. The hotel staff also made themselves available to clarify any questions that customers may have regarding their stay, the hotel or the region, prior to their arrival. This information enabled hotels to provide a better and customized service to customers, also enhancing the quality of service.

The system identified a high number of predicted cancellations. Since the hotels did not have sufficient resources to contact all customers, hoteliers defined selection criteria for which bookings were to be contacted:

1. The arrival date should be three days in advance of the current date, at a minimum;
2. The booking should be made at a reasonable price or yield high room revenue;
3. The costumer had to be directly contactable (e.g., extranet contacts or direct emails). Note that this criterion excluded any customers who were traveling with traditional travel agencies or other partners not disclosing direct contact with their customers (e.g., Hotelbeds).
4. The costumer's nationality and language were identifiable and in which there was some proficiency by some of the hotel staff. Therefore, hotels only contacted customers who spoke Portuguese, Spanish, German, English or French.
5. Only bookings classified as likely to cancel during at least 50% of the time the booking had been processed by the model (MF) should be chosen. However, this criterion was not mandatory: if resources were available, bookings with lower frequencies representing a high revenue would be contacted as well.

The majority of the contacts was made via personalized direct emails or their original booking platform (e.g., Booking.com extranet or Expedia.com extranet). Using templates for each language, texts were always personalized for each customer.

5.3 Evaluation

In this section, both quantitative results and the qualitative results are presented and discussed. The rationale behind the qualitative results is encompassed in the interviews with the hotel chain revenue management team during the time of the system's deployment. These conversations exposed interesting results that cannot be quantitatively captured.

5.3.1 Quantitative results

The proposed approach shows that the capacity of the system to continuously learn with the daily incorporation of new bookings—with changes to existing bookings and with the outcome of previous predictions—and the ability to automatically build a new model every day produced a system that achieved good quantitative results.

The chosen "Champion-challenger" strategy showed that the system required a relatively short time to stabilize. In the case of R1, the system commuted to the challenger model only twice within the first two weeks of deployment. Similarly, for C1, the system changed four times in the first four weeks of deployment. Since then, the champion model has been consistent. This stability does not imply that the model would not change again but implies that the system only changes after a proven performance. This finding can be explained by the criteria specifications for the challenger model to be selected, requiring the challenger to demonstrate superior performance when compared with the performance of the champion model. The criteria ensure that a challenger model that performed very well on a particular day is not promptly selected.

From the perspective of standard machine learning performance metrics, since models were built and assessed daily, it is difficult to present results for the entire assessment period. Because daily results were very similar, only the performance metrics for the last day are presented (Table 13). As expected, the results are inferior to those reported in the exploratory models but in line with final models (Model 1 and 2). Compared to exploratory models, are less prone to overfitting, more robust, and do not exhibit problems of over-classification for future arrivals. On August 31st, 2017, the percentage of future arrivals identified as likely to cancel was 26.4% for C1, and 18.6% for R1, which is consistent with hotels previous cancellation rates (Figure 48). Similarly, differences among hotels' cancellation rates are also present in the models' performance metrics, which consistently present superior values for C1.

Table 13 - Performance metrics on August 31st, 2017

Hotel	Dataset	Accuracy	Precision	F1Score	AUC	Sensitivity	Specificity
C1	Train	0.8701	0.8849	0.8460	0.9438	0.8103	0.9171
	Test	0.8563	0.8731	0.8274	0.9276	0.7862	0.9110
R1	Train	0.8646	0.8484	0.7410	0.9227	0.6577	0.9510
	Test	0.8486	0.8205	0.7016	0.8864	0.6128	0.9452

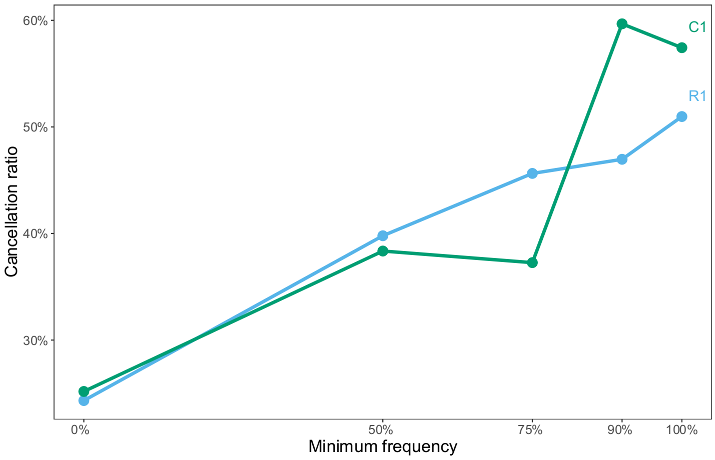
A/B testing also presented compelling results. For arrivals expected between June 2017 and August 2017 (excluding bookings canceled prior to the model deployment, that is, April 2017), the number of bookings on which hotels acted to avoid cancellations was rather low (4.8% for C1 and 5.4% for R1). The percentage of canceled bookings in group "A" (the group kept from users) is 0.6% higher than the results for group "B" (Table 14). This finding translates into a relative decrease in group "B" cancellations of 2.0% for C1 and 2.5% for R1. Nevertheless, it should be noted that the differences are not sufficient to consider the results as statistically significant. Cohen's h size effect (Cohen 1988), i.e., the difference in the cancellation rate, would have to exceed 5.5% for C1 and exceed 7.9% for R1 (at a significance level of 0.05, using a power of test[8] of 0.80). The Chi-square test of independence also shows that this difference is not statistically significant for both, with a value of x^2(1)=0.234, p=0.629 for C1 and x^2(1)=0.144 and p=0.705 for R1.

Table 14 - A/B testing effective cancellation summary

Hotel	Group	Canceled	Not canceled	Total	% Canceled	Actions	% Actions
C1	A	1 043	3 060	4 103	25.4%	N/A	N/A
	B	1 025	3 086	4 111	24.9%	196	4.8%
R1	A	486	1 489	1 975	24.6%	N/A	N/A
	B	483	1 526	2 009	24.0%	109	5.4%

Assessing the system using the MF ratio confirms the system's predictions precision. As depicted in Figure 53, an MF decrease is followed by a decrease in the cancellation ratio. The cancellation ratio for bookings with an MF of 100%, or in other words, bookings that were predicted as likely to cancel every time they were processed, was 57.4% for C1 and 50.1% for R1. These values decrease to 38.4% for C1 and 39.8% for R1 with bookings that were predicted as likely to cancel at least 50% of the times that they were processed (MF≥50%). These values contrast with the total cancellation ratio of 25.2% for MF≥0% with C1 and 24.3% with R1.

Figure 53 - Cancellation ratio by minimum frequency



Note: MF threshold levels were selected based on the users' criteria to select the bookings to contact. Mostly, users only selected bookings with an MF equal to or greater than 50%.

[8] Percentage of the minimum of time that the minimum effect size will be detected (assuming it exists).

Note that the cancellation ratio could be higher if hotels had not contacted some of the bookings to avoid cancellation. In fact, considering the low number of bookings acted on to prevent cancellations in relation to the total number of bookings that were predicted as likely to cancel (Table 14), the actions had a significant impact on avoiding cancellations. The analysis of the “B” groups that were displayed to the hotels shows a substantial difference in terms of the cancellation rates between the bookings were no actions were taken and bookings that were acted upon (Table 15). The difference is of 18.1 percentage points for C1 for all “B” group bookings with MF ≥ 0%, translating into a relative decrease in cancellations of 70%. For R1, the difference is 13.8 percentage points for R1, which translates into a relative decrease in cancellations of 56%. A Chi-square test of independence confirms that the difference is statistically significant for both hotels: C1: x^2 (1)=31.873, p<0.001; R1: x^2 (1)=9.978, p=0.002. For “B” group with MF≥50%, the differences are even more substantial: 37.8 percentage points for C1 and 37.1 percentage points for R1 for the cancellation ratio, which corresponds to relative decreases in cancellations of 84% for C1 and 82% for R1. A Chi-square test of independence confirms that this difference is statistically significant for both: C1: x^2(1)=58.373, p<0.001; R1: x^2(1)=33.609, p<0.001.

The effect of contacting customers can be compared if bookings for which customers were contacted and bookings for which customers were not contacted are measured. For bookings with an MF≥50%, not contacting the guest entails a cancellation enhancer factor at a magnitude of 10.0 for C1, and a magnitude of 9.3 for R1, with 95% CIs [5.26, 21.74] and [4.20, 24.83], respectively. The lower cancellation rate overall bookings contacted by hotels, independently of their prediction as likely to cancel (MF≥0%), indicates that contacting customers may reduce the number of cancellations. Because contacting all customers requires resources that are unavailable most of the time, these results highlight the importance of having a booking cancellation prediction model to identify bookings to reduce the resources required to contact customers.

Table 15 - "B" group cancellation results summary

Hotel	Action	MF≥0% (all bookings)			MF≥50%		
		Canceled	Not canceled	% Canceled	Canceled	Not canceled	% Canceled
C1	No	1 010	2 905	25.8%	269	325	45.3%
	Yes	15	181	7.7%	9	111	7.5%
R1	No	471	1 429	24.8%	125	153	45.0%
	Yes	12	97	11.0%	6	70	7.9%

From a financial perspective, despite the low number of contacted customers, the analysis of the results emphasizes the impact to prevent cancellation of bookings identified as likely to cancel.

Considering the proportion of bookings where actions to prevent cancellations were taken and did not effectively cancel in relation to those with no actions taken, the room revenue that has not lost to cancellations amounts to € 22,144.77 for C1 and € 16,680.97 for R1. For both, the actions taken prevented a total revenue loss of € 38,825.75, which corresponds to a monthly average of € 12,941.91 of room revenue that is not lost to cancellations during the three months of the system's deployment. Some of this value would not have been lost even if cancellations occurred since hotels would eventually re-sell some of the rooms' nights. Cancellations increase uncertainty and prevent hotels' revenue management teams to increase prices, confirming the positive impact on the hotel business performance of contacting customers of bookings that are identified as likely to cancel.

Another interesting aspect is the fact that some customers who were contacted replied on the same day or the following day with an effective cancelation. This finding may not be negative since hotels can immediately put the rooms for sale again.

5.3.2 Qualitative results

From the periodic interviews with the hotel chain revenue management team and the final interview, four important considerations were highlighted.

First, users suggested that the system should be fully integrated with the PMS or should be able to display each booking's complete details. Users indicated that this requirement could expedite the time required to identify the details of each booking that was predicted as likely to cancel. This situation also limits the total number of customers of that they manage to contact about their bookings. Note that this limitation only existed because of the research nature of this project.

Second, hotels recognized that they seldom took advantage of the "net demand" as an indicator in their demand-management decisions and acknowledged their resistance to change, instead of a lack of confidence in the system, as the main reason. In situations in which the hotel was overbooked or situations that required decisions for short-term dates, they considered the system "net demand" metric to decide whether to open or close sales at certain time. As an example, the C1 team mentioned that one day, at approximately 06:00 PM and with the hotel fully booked for the night, they decided to accept two walk-ins[9] because the system identified that four of the bookings still without to check-in were identified by the system as likely to cancel. Half of these four bookings canceled.

Third, hotel users recognize that the system may have a positive impact on the hotel's social reputation because most customers who were contacted engaged in conversation with the hotel staff, thanked them for their concern and allowing the hotel to provide them with better service.

[9] "Walk-in" is a term used in hotel revenue management to describe customers that arrived at the property seeking a room without an advanced reservation.

Last, all users positively answered when asked if they would continue to use the system if it was made available as a permanent tool.

5.4 Discussion

From a scientific standpoint, this chapter addressed several of the roles that predictive analytics can assume in scientific research, namely, the development of new measures, the improvement of existing models, and the relevance and predictability assessment. The system's need allowed for the definition of a new measurement—MF—for evaluating the performance of binary classification problems where observation characteristics are unstable or where the outcome of the prediction is affected by time. The development and deployment of the system demonstrated how the data-splitting method and domain knowledge in feature engineering are paramount for machine learning modeling and its influence on the improvement of existing prediction models. The development and deployment of the models in a prototype tested under real-world conditions enabled the assessment of the system's relevance and predictability. This demonstration evidences the benefits of machine learning for business information systems, as advocated by several authors. However, the benefits for applied research remained ambiguous.

Another point that distinguishes the development of the system is the use of open-source tools such as Linux, R, and Hadoop. The system's performance and results proved the adequacy and usefulness of these tools for the problem of booking cancellation prediction. The Linux Hadoop/Spark cluster running R Server enabled the modeling process to be distributed through different cluster machines, taking advantage of the available computational power and the powerful XGboost tree boosting machine learning algorithm. The results validated the value of the system architecture design for running an automated machine learning system, the daily incorporation of new data, and employing previous prediction errors and hits to improve continuously.

From a business standpoint, the system presents significant results. First, it shows similar final results for the different hotels: *Accuracy* greater than 0.84, *Precision* greater than 0.82, and *AUC* greater than 0.88. Second, the bookings cancellation ratio in the ones predicted as likely to cancel in at least half of the days' processed (MF≥50%) attained 38.4% for C1 and 39.8% for R1. These results exceed the cancellation ratio of all bookings (MF≥0%): 25.2% for C1 and 24.3% for R1. The results are even more noteworthy if some of the bookings that were identified as likely to cancel were contacted had their potential cancellation was reverted. Quantitative results stress the satisfactory level of precision of the models. Third, despite the difficulties associated with contacting customers prior to their arrival (including the costs associated with the contact), the identification of possible cancellations enables acting to prevent cancellations at a limited cost. The decrease in the number of actual cancellations on bookings for which customers were contacted, a total in excess of 37 percentage points, corresponds to a relative cancellation decrease of 83% for C1 and 82% for R1. These findings indicate that the actions taken prevented cancellations whose total revenue is in the order of approximately € 39,000.00. Although not all of the future bookings identified as likely to cancel can be contacted, results indicate that an increase in the number of contacted customers should prevent additional cancellations and revenue loss.

Scalability is a common problem in machine learning problems (Domingos 2012). The presented results obtained using with only one source of data - PMS, shows that it is possible to build predictive analytics systems without the need of vast resources and multiple data sources.

Overall, this prototype highlights how an automated machine learning system, designed in accordance to DSR to address an unsolved problem in a unique and innovative manner, can be forged, implemented and having a measurable impact on business. The benefits for revenue management in service-based industries of exploiting mathematical and forecast models to take advantage of technology and the data available are confirmed. In addition, by showing its efficacy and suitability, the system establishes how bookings cancellation prediction models can be integrated into RMS, thus answering affirmatively to RQ3.

5.5 Summary

A prototype was designed and deployed in two hotels to study how an RMS component that predicts hotel bookings cancellation could be built and implemented. The prototype consisted of a machine learning model, based on final Models 1 and 2. The system ran in an automated form, collecting and processing daily hotels' PMS data to predict bookings cancellation.

The system prototype employed the "champion-challenger" approach and a mechanism for learning from previous hits and misses, so to continuously improve the models' performance and adapt to changes in cancellations patterns.

Linux Hadoop/Spark cluster running R Server proved to be a good choice for taking advantage of distributing computing and of the XGboost tree boosting machine learning algorithm.

Performance metrics, together with hotel users' assessment, confirmed the system design and architecture to be suitable for a tool in revenue management. The use of the system allowed hotels' staff not only to make better demand-management decisions based on their net demand forecast but also to decrease cancellations by contacting some of the bookings predicted as "likely to cancel".

The prototype also addressed several of the roles that, as previously introduced, Shmueli, Koppius (2011) recognize that predictive analytics can assume in research. In this case, the improvement of existing models, the relevance and predictability assessment, and the development of new measures. Of note is the latter, since it led to the development of MF – a measure introduced to evaluate the performance of binary classification problems where observation characteristics are unstable or where the outcome of the prediction is affected by time.

In short, besides its impact in predictive analytics research, the system development, deployment, and assessment confirmed that the deployment framework designed in the exploratory models could be implemented and could translate in benefits for hotels, thus answering RQ3 affirmatively.

6 CONCLUSION

Predicting booking cancellations is of the foremost importance for hotel revenue management, not only to estimate net demand but also to understand what drives cancellations. Nonetheless, booking's cancellation prediction, as Chapter 2 demonstrates, is still an understudied topic, particularly in the hotel industry. The availability of higher computational power at lower costs and advances in data science disciplines, such as data mining, machine learning, and data visualization, seems to have enhanced the interest for the topic, reflected in a growing number of publications since 2008. Even taking into account other travel and tourism-related industries, only a small number of publications explored the potential of using advanced machine learning classification algorithms and booking detailed data to predict booking cancellations at a disaggregated level. This dissertation aims to fill this gap and make use of data science tools and capabilities to develop hotel bookings' cancellation prediction models. In addition, this dissertation intends to disclose which factors drive cancellations, how cancellation models could be deployed, and how models impacted business operations.

Section 6.1 summarizes the answers to the three research questions and links them with the chapters where the answers were addressed. Section 6.2 provides an overall description of the research findings and contributions. The chapter ends with a discussion of research limitations and opportunities for further research in Section 6.3.

6.1 Answers to the research questions

Section 2.2 (Chapter 2) summarizes the state of the art in hotel bookings' cancellation prediction modeling, helping to define the focus, methodology, analysis and reporting format of the dissertation through the formulation of three main research questions. Chapters 3, 4, and 5 provide the answers for each one of the research questions. The current section intends to present a summary of those answers, how they were reached and where in the dissertation are they presented.

RQ1. Could a booking's cancellation prediction model that uses PMS data display better results than a model that uses PNR data?

As initially hinted by the modeling results in Chapter 3 and later confirmed by the results presented in Chapters 4 and 5, the use of features extracted or engineered from PMS data proves to be adequate to build hotel bookings' cancellation prediction models than data in a pre-established format. Unlike PNR data which follows a base format originally designed for the airline industry, PMS data only depends on the PMS's database structure. Because PMS hold most of the data related to hotel bookings, the level of detail and the features to include in the modeling dataset depend only on the modeler domain expertise and PMS database structure knowledge.

RQ2. Could this model be improved with the inclusion of data from additional sources?

Contrary to plausible expectations, supported by some authors, about the potential of big data in forecasting (see Chapters 1 and 2), big data in the form of features from a variety of sources, gathered in enormous quantities, does not improve the predictive models' performance. As exposed in Chapter 4, though some features from other non-PMS data sources, for some hotels, had explanatory value, they did not carry any predictive value. Due to the amount of resources that are required to collect and process data from other sources, such as social reputation data or online prices/inventory, the use of non-PMS data in bookings cancellation prediction models should only be considered if features created from those sources could have significant predictive power.

RQ3. Can such a model be integrated into an hotel RMS?

To answer this question, a framework for the integration of the booking cancellation prediction models into RMS has been proposed in Chapter 3. This framework has been tested in a real production environment through the development and deployment of a prototype in two different types of hotel. The results discussed in Chapter 5 confirmed, not only the framework design adequacy but also the impact bookings' cancellation prediction models can have on business operations.

6.2 Contributions and implications

Prediction models for booking's cancellation classification like the ones that have been presented enable hotels to determine their true demand, not only at a global level but also in a disaggregated form, by room type, market segment, distribution channel, among other factors. Moreover, by showing what drives cancellations, prediction models allow revenue managers to adjust overbooking tactics and cancellation policies. Hence, an hotel could present less restrictive policies to its customers predicted as unlikely to cancel, while introducing more restrictive policies for customers predicted as more likely to cancel. The application of less restrictive cancellation policies has the potential to increase the number of bookings by not applying restrictive policies indiscriminately, and increase revenue by reducing the need to make discounts and thus decreasing the number of bookings with restrictive cancellation policies. Additionally, if overbooking is employed selectively, hotels could decrease losses related to reallocation costs of immediate and future revenue from “walked” customers. Furthermore, booking cancellation classification models allow hotels to intervene prior to check-in and act to prevent cancellations, thus reducing the loss of revenue associated to cancellations and, at the same time, reducing revenue manager’s temptation to offer discounts or special offers to make-up for cancellations.

The fact that classification models built with detailed booking data can determine each booking’s cancellation probability does not mean that the use of historical data and the building of regression models can be discarded in hotel revenue management, quite the opposite. If the objective is to forecast long-term net demand, especially for a period longer than the average lead time, then, regression models should be built using historical data and appropriate algorithms, since only a few bookings may exist on-the-books.

The performance results described highlighting the importance that machine learning can have for hospitality management, particularly in revenue management. Estimation and forecasting is one of the essential processes of revenue management and machine learning can help managers improve results, with superior accuracy, in a more timely manner, and, above all, in a more pragmatic way, not so dependent on personal guesses or speculations. Overall, by reducing uncertainty, booking cancellation prediction models allow revenue managers to make better structural, pricing or quantity demand-management decisions.

The main theoretical, methodological and practical contributions of this dissertation, together with the main implications are summarized below.

- Theory:
    - A critical analysis of the state of the art in terms of bookings cancellation prediction/forecast for the travel industry and, in particular, in the hotel industry;
    - Demonstration of the value of highly detailed hotel-specific data (PMS data) in detriment of the more usually adopted PNR format for predicting the probability of a hotel booking being canceled with high accuracy;
    - Discovery of which features have higher predictive power for understanding cancellation drivers and a demonstration of how this knowledge could be employed in the formulation of cancellation policies;
    - Deflation of the impact of big data for revenue management forecasting problems and forewarn towards the relation between needed resources versus benefits;
    - Draw of the attention to how competitive sets are defined and to the possible benefits of using dynamic cancellation policies, identifying these as topics where further research is needed;
    - Definition of a new measure – MF – for evaluating the performance of binary classification problems when observations characteristics are unstable or when the outcome of the prediction is affected by time;

    - The development and deployment of a system prototype that confirmed the added value of predictive analytics in scientific research as an instrument for the design of evaluation measurements (MF measure), the improvement of existing models (changes on field test models when comparing with Model 1 and 2) and to assess the relevance and predictability of models.
- Methods:
    - Evidence of how to do semiautomated literature reviews. Source code and data released in a paper presented in the 2018 edition of the TMS conference and selected for publication in the Tourism & Management Studies journal;
    - Identification of which input and which engineered features are necessary to build models for booking cancellation prediction;
    - Methodology for the collection and splitting of PMS' features to taper off leakage and overfitting problems;
    - Identification of data sources and detailed description of data collection from multiple sources, namely: currencies exchange values, events in hotels' regions, holidays per country, stock exchange indexes, and weather forecasts;
    - Exhibition of the benefits of data science tools and capabilities, such as data visualization, statistics and data mining for the apprehension of patterns and tendencies in hotel booking data;
    - Demonstration of continuously automated machine learning systems that learn from their predictions together with new data, and of how can they can be built, parameterized and deployed.
- Practice:
    - Demonstration of how contacting customers prior to their arrival reduces cancellations, preventing revenue loss and simultaneously helping revenue managers to be more assertive in their demand-management decisions;
    - Show that with detailed net demand forecasts, hotels can make more informed overbooking decisions, thus helping to mitigate reallocation, compensation and reputation costs associated with overbooking.

From the point of view of DSR, this dissertation makes valuable contributions in the three types of possible DSR research contributions: novelty, generality, and significance. Being representation fidelity and "implementability", together with a clear demonstration of improvement of the business problem, the criteria for accessing contribution in the scope of DSR (Hevner, March, Park, Ram 2004), the development and implementation of the prototype definitely demonstrated this dissertation contribution. The design artifact (prototype) demonstrated how an unsolved problem could be dealt with, how existing methods could be combined and extended to solve the business problem, and how the association of different methods improved the knowledge of the business problem and fostered the development of a new measure that has the potential of being applied to other types of problems.

6.3 Limitations and future research

Like any other research, this dissertation has limitations, some of which are an opportunity for further research.

Machine learning models' product is a very complex prediction equation that does not allow for the models to be depicted. Nevertheless, researchers can follow the steps described in this dissertation to replicate the models here presented.

Regarding the model's construction, despite the inclusion of data from multiple sources being advocated as a way to improve forecasting performance (McGuire 2017; Pan, Yang 2017a; Talluri, Van Ryzin 2005; Wang, Yoonjoung Heo, Schwartz, Legohérel, Specklin 2015; Zhang, Shu, Ji, Wang 2015), results contradict that claim. Although some features from non-PMS data sources have shown explanatory importance in terms of cancellations, predictive models did not corroborate that importance. This result could be due to the lack of relevance in terms of booking cancellations of the data sources employed, or due to the lack of predictive importance of the features engineered. Consequently, future research could employ features from additional data sources or engineer different features from the same data sources. For example, in terms of PMS data and for hotels that work with recurring groups from the same travel operators (common in many city hotels), a feature that could capture those groups "wash"[10] may prove to have predictive importance. In terms of social reputation data, instead or in complement to the hotel overall quantitative ratings, a feature that represents reviews' textual component sentiment polarity could be used.

Features engineered from the hotels' competitive sets' social reputation and online prices/inventory did not show any predictive importance, i.e., better social rating or better prices of competitors did not influence cancellations. This raises questions on how competitive sets are defined and about the effectiveness of using competitive sets for some kinds of problems. Are today's competitive sets helpful in the hospitality industry? For some types of travelers, this could be questionable. For someone deciding on whether to book holidays in Portugal, Spain, or Cyprus and making multiple hotel reservations in these countries, the hotel's competitors will be outside of its local hotel competitors set. The same applies for someone deciding on whether to book a weekend break in Lisbon, Barcelona, or London. Therefore, demand forecast research should consider the use of other data sources, like on-the-books sales data, or demand forecast data for competing regions or destinations. However, those data sources could be difficult to obtain. To overcome this, heuristics could be created from other data sources like airport passenger traffic forecasts, or cruise departures and arrivals. These data sources may be employed to complement the hotel's competitive set data.

[10] The term "wash" is used in hotel revenue management to quantify the difference between the quantity of rooms a group pledges to occupy and the rooms it effectively occupies.

This dissertation demonstrated that it is possible to understand the predictive importance of each feature in terms of cancellation and that this importance differs per hotel. Future research could explore this knowledge in order to develop models that determine dynamic cancellation policies. Those models could be employed in hotel/brand websites and other online distributors' websites, in order to adjust cancellation policies according to the details of each booking search and according to the cancellation probability.

Regarding the system prototype, limitations were essentially related to the reduced resources that hotels could allocate to the project (amount and time) and the limitations of being a research project. The first limitation was the shortening of the test period by one month. The second limitation was the difficulty to collect the number of customers who responded to the hotels' contact. This metric could have been interesting for measuring the effective reach of the customers' contact. However, due to the multiplicity of channels that a customer can use to reach a hotel and the many different persons/departments who can handle the contact, registering this process was not possible. The hotels' revenue management team estimates this number to be very low, probably less than 10%.

Two additional limitations, which were imposed by research requirements, also contributed to the low number of contacted bookings. The fact that the system was designed to include A/B testing did not allow hotel users to obtain the details of bookings in the "A" group limited the number of bookings hotels could contact. Also, the amount of time invested in the selection of the bookings to contact and the time required to obtain the contacts of these bookings, because it required the consultation of booking details in the hotels' PMS, also contributed for the relatively low number of bookings contacted. In a real production system, all bookings identified by the system as likely to cancel could be contacted, allowing hotels to contact a larger number of customers more efficiently.

Approximately two years of data were available for training, but the modeling dataset did not include features that could explicitly capture seasonality, such as the *DayOfYear*, subsequently implemented in final models 1 and 2. The hospitality industry, especially for resort hotels, is an industry where seasonality has an important influence on business. The use of data from a wider timespan with the inclusion of time/season-specific features has the potential to enable the development of models with better performance.

The system itself has the potential to generate new features that may have an important role in improving models' performance. Since bookings that were acted upon are canceled less frequently than bookings in which no action was taken, a feature with the indication of which category of action was taken, if any, is expectable to improve model performance. Additionally, recording the actions taken for each booking to avoid cancellation (e.g., offering a room upgrade or asking about the bed type preference) has potential use in another machine learning model, capable of recommending which actions should be executed for each of the bookings predicted as likely to cancel. This finding can prompt the development of a fully automated system. A system that not only can predict the bookings cancellation outcome but can also select which customers to contact, make initial contact and engage in discussion with the customer via a chatbot, only requiring human intervention in the aspects of the discussion where the system is not prepared to answer.

REFERENCES

ABBOTT, Dean, 2014. Applied predictive analytics: Principles and techniques for the professional data analyst. Indianapolis, IN, USA: Wiley. ISBN 978-1-118-72796-6.

ABE, Naoki, ZADROZNY, Bianca and LANGFORD, John, 2004. An iterative method for multi-class cost-sensitive learning. In: Proceedings of the tenth ACN SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA. August 2004. p. 3–11.

AL SALEEM, Abdul and AL-JUBOORI, Noorya, 2013. Factors affecting hotels occupancy rate (An empirical study on some hotels in Amman). Interdisciplinary Journal of Comtemporary Research in Business [online]. October 2013. Vol. 5, no. 6. [Viewed 4 October 2015]. Available from: http://journal-archieves36.webs.com/142-159.pdf

ALI, Nauman Bin and USMAN, Muhammad, 2018. Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy. Information and Software Technology. July 2018. Vol. 99, p. 133–147. DOI 10.1016/j.infsof.2018.02.002.

ANDERSON, Chris K., 2012. The impact of social media on lodging performance. Cornell Hospitality Report. 2012. Vol. 12, no. 15, p. 4–11.

ANTONIO, Nuno, ALMEIDA, Ana and NUNES, Luis, 2017. Predicting hotel booking cancellation to decrease uncertainty and increase revenue. Tourism & Management Studies. 2017. Vol. 13, no. 2, p. 25–39. DOI 10.18089/tms.2017.13203.

APILAYER, [no date]. apilayer - Automate what should be automated. Apilayer [online]. [Viewed 23 August 2018]. Available from: https://apilayer.com/

ARUN, R., SURESH, V., MADHAVAN, C. E. Veni and MURTHY, M. N. Narasimha, 2010. On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In: Advances in Knowledge Discovery and Data Mining [online]. Springer, Berlin, Heidelberg. 21 June 2010. p. 391–402. [Viewed 27 October 2017]. Lecture Notes in Computer Science. ISBN 978-3-642-13656-6. Available from: https://link.springer.com/chapter/10.1007/978-3-642-13657-3_43

AZADEH, Shadi Sharif, LABIB, Richard and SAVARD, Giles, 2013. Railway demand forecasting in revenue management using neural networks. International Journal of Revenue Management. 2013. Vol. 7, no. 1, p. 18. DOI 10.1504/IJRM.2013.053358.

AZADEH, Shadi Sharif, 2013. Demand forecasting in revenue management systems [online]. École Polytechniqye de Montreál, Canada. [Viewed 24 July 2018]. Available from: https://publications.polymtl.ca/1216/1/2013_ShadiSharif_Azadeh.pdf

AZEVEDO, Ana and SANTOS, Manuel F., 2008. KDD, semma and CRISP-DM: A parallel overview. In: MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008. 2008. p. 182–185.

BARNES, Jeff, 2015. Azure Machine Learning. Redmond, WA: Microsoft Press. Microsoft Azure Essentials. ISBN 978-0-7356-9817-8.

BATRINCA, Bogdan and TRELEAVEN, Philip C., 2015. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY. February 2015. Vol. 30, no. 1, p. 89–116. DOI 10.1007/s00146-014-0549-4.

BENÍTEZ-AURIOLES, Beatriz, 2018. Why are flexible booking policies priced negatively? Tourism Management. August 2018. Vol. 67, p. 312–325. DOI 10.1016/j.tourman.2018.02.008.

BERGSTRA, James S, BARDENET, Rémi, BENGIO, Yoshua and KÉGL, Balázs, 2011. Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems 2011. Granada, Spain. 2011. p. 9.

BJØRKELUND, Eivind, BURNETT, Thomas H. and NØRVAG, Kjetil, 2012. A study of opinion mining and visualization of hotel reviews. In: Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services [online]. ACM. 2012. p. 229–238. [Viewed 7 November 2015]. Available from: http://dl.acm.org/citation.cfm?id=2428773

BLEI, David M., NG, Andrew Y. and JORDAN, Michael I., 2003. Latent dirichlet allocation. Journal of machine Learning research. 2003. Vol. 3, no. Jan, p. 993–1022.

BOSCH, Olav ten, 2017. An introduction to web scraping, IT and legal aspects. European Comission [online]. 2017. [Viewed 26 August 2018]. Available from: https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2017%20ESTP%20PROGRAMME/45.%20Automated%20collection%20of%20online%20prices_%20sources%2C%20tools%20and%20methodological%20aspects%2C%2023%20%E2%80%93%2026%20October%202017%20-%20Organiser_%20EXPERTISE%20FRANCE/20170919%20ESTP%20Prices%20-%206%20-%20Introduction%20IT%20and%20Legal.pdf

BRAGGE, Johanna, RELANDER, Sami, SUNIKKA, Anne and MANNONEN, Petri, 2007. Enriching literature reviews with computer-assisted research mining. Case: profiling group support systems research. In: [online]. IEEE. 2007. p. 243a-243a. [Viewed 6 May 2018]. Available from: http://ieeexplore.ieee.org/document/4076874/

BRAUN, Michael T., KULJANIN, Goran and DESHON, Richard P., 2018. Special considerations for the acquisition and wrangling of big data. Organizational Research Methods. July 2018. Vol. 21, no. 3, p. 633–659. DOI 10.1177/1094428117690235.

CALHEIROS, Ana Catarina, MORO, Sérgio and RITA, Paulo, 2017. Sentiment classification of consumer-generated online reviews using topic modeling. Journal of Hospitality Marketing & Management. 27 March 2017. Vol. 0, no. 0, p. 1–19. DOI 10.1080/19368623.2017.1310075.

CANTALLOPS, Antonio S. and SALVI, Fabiana, 2014. New consumer behavior: A review of research on eWOM and hotels. International Journal of Hospitality Management. 2014. Vol. 36, p. 41–51. DOI 10.1016/j.ijhm.2013.08.007.

CHANDRASHEKAR, Girish and SAHIN, Ferat, 2014. A survey on feature selection methods. Computers & Electrical Engineering. 1 January 2014. Vol. 40, no. 1, p. 16–28. DOI 10.1016/j.compeleceng.2013.11.024.

CHAPMAN, Pete, CLINTON, Julian, KERBER, Randy, KHABAZA, Thomas, REINARTZ, Thomas, SHEARER, Colin and WIRTH, Rudiger, 2000. CRISP-DM 1.0: Step-by-step data mining guide. The Modeling Agency [online]. 2000. [Viewed 10 September 2015]. Available from: https://the-modeling-agency.com/crisp-dm.pdf

CHEN, Chiang-Ming, TSAI, Yi-Chun and CHIU, Hsien-Hung, 2017. The decision-making process of and the decisive factors in accommodation choice. Current Issues in Tourism. 25 January 2017. Vol. 20, no. 2, p. 111–119. DOI 10.1080/13683500.2015.1087476.

CHEN, Chih-Chien, SCHWARTZ, Zvi and VARGAS, Patrick, 2011. The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. International Journal of Hospitality Management. March 2011. Vol. 30, no. 1, p. 129–135. DOI 10.1016/j.ijhm.2010.03.010.

CHEN, Chih-Chien and XIE, Karen (Lijia), 2013. Differentiation of cancellation policies in the U.S. hotel industry. International Journal of Hospitality Management. September 2013. Vol. 34, p. 66–72.

CHEN, Chih-Chien, 2016. Cancellation policies in the hotel, airline and restaurant industries. Journal of Revenue and Pricing Management. 25 March 2016. Vol. 15, no. 3–4, p. 270–275. DOI 10.1057/rpm.2016.9.

CHEN, Tianqi and GUESTRIN, Carlos, 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [online]. ACM. 2016. p. 785–794. [Viewed 22 January 2017]. Available from: http://dl.acm.org/citation.cfm?id=2939785

CHIANG, Wen-Chyuan, CHEN, Jason CH and XU, Xiaojing, 2007. An overview of research on revenue management: current issues and future research. International Journal of Revenue Management. 2007. Vol. 1, no. 1, p. 97–128. DOI 10.1504/IJRM.2007.011196.

CIRILLO, Cinzia, BASTIN, Fabian and HETRAKUL, Pratt, 2018. Dynamic discrete choice model for railway ticket cancellation and exchange decisions. Transportation Research Part E: Logistics and Transportation Review. February 2018. Vol. 110, p. 137–146. DOI 10.1016/j.tre.2017.12.004.

CLEMENTS, Michael and HENDRY, David, 1998. Forecasting economic time series. Cambridge University Press. ISBN 978-0-521-63480-9.

CLEVEN, Anne, GUBLER, Philipp and HÜNER, Kai M., 2009. Design alternatives for the evaluation of Design Science research artifacts. In: Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology [online].

New York, NY, USA: ACM. 2009. p. 19:1–19:8. [Viewed 26 August 2014]. DESRIST '09. ISBN 978-1-60558-408-9. Available from: http://doi.acm.org/10.1145/1555619.1555645

COHEN, Jacob, 1988. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates. ISBN 978-0-8058-0283-2. HA29 .C66 1988

COVER, Thomas M. and THOMAS, Joy A., 1991. Elements of information theory [online]. New York, NY: John Wiley & Sons, Inc. [Viewed 17 August 2018]. ISBN ISBN 0-471-20061-1. Available from: http://epubs.siam.org/doi/10.1137/1036124

CROSS, Dax, 2016. A history of revenue management and the advent of next-generation RM. Journal of Revenue and Pricing Management. 1 July 2016. Vol. 15, no. 3–4, p. 293–298. DOI 10.1057/rpm.2016.5.

DAY, Jonathon, CHIN, Natalie, SYDNOR, Sandra and CHERKAUER, Keith, 2013. Weather, climate, and tourism performance: A quantitative analysis. Tourism Management Perspectives. January 2013. Vol. 5, p. 51–56. DOI 10.1016/j.tmp.2012.11.001.

DEKAY, Frederick, YATES, Barbara and TOH, Rex S., 2004. Non-performance penalties in the hotel industry. International Journal of Hospitality Management. September 2004. Vol. 23, no. 3, p. 273–286.

DELEN, Dursun and CROSSLAND, Martin D., 2008. Seeding the survey and analysis of research literature with text mining. Expert Systems with Applications. April 2008. Vol. 34, no. 3, p. 1707–1720. DOI 10.1016/j.eswa.2007.01.035.

DENIZCI GUILLET, Basak and MOHAMMED, Ibrahim, 2015. Revenue management research in hospitality and tourism: A critical review of current literature and suggestions for future research. International Journal of Contemporary Hospitality Management. 11 May 2015. Vol. 27, no. 4, p. 526–560. DOI 10.1108/IJCHM-06-2014-0295.

DHAR, Vasant, 2013. Data science and prediction. Communications of the ACM. 1 December 2013. Vol. 56, no. 12, p. 64–73.

DOMINGOS, Pedro, 2012. A few useful things to know about machine learning. Communications of the ACM. 2012. Vol. 55, no. 10, p. 78–87.

ENZ, Cathy A., CANINA, Linda and LOMANNO, Mark, 2009. Competitive pricing decisions in uncertain times. Cornell Hospitality Quarterly. August 2009. Vol. 50, no. 3, p. 325–341. DOI 10.1177/1938965509338550.

ENZ, Cathy A., CANINA, Linda and WALSH, Kate, 2001. Hotel-industry averages: An inaccurate tool for measuring performance. Cornell Hotel and Restaurant Administration Quarterly. 2001. Vol. 42, no. 6, p. 22–32. DOI 10.1016/S0010-8804(01)81005-3.

EUROPEAN COMMISSION (ed.), 2014. Study on online consumer reviews in the hotel sector. Final report. [online]. 2014. European Union. [Viewed 28 June 2016]. Available from: http://rpaltd.co.uk/uploads/report_files/hotel-reviews.pdf

EYSENBACH, Gunther, TUISCHE, Jens and DIEPGEN, Thomas L., 2001. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. Medical Informatics and the Internet in Medicine. January 2001. Vol. 26, no. 3, p. 203–218. DOI 10.1080/14639230110075459.

FEINERER, Ingo and HORNIK, Kurt, 2017. tm: Text mining package [online]. Available from: https://CRAN.R-project.org/package=tm

FELLOWS, Ian, 2014. wordcloud: Word clouds [online]. Available from: https://CRAN.R-project.org/package=wordcloud

FENG, L., CHIAM, Y. K. and LO, S. K., 2017. Text-mining techniques and tools for systematic literature reviews: A systematic literature review. In: 2017 24th Asia-Pacific Software Engineering Conference (APSEC). December 2017. p. 41–50.

FERNÁNDEZ-DELGADO, Manuel, CERNADAS, Eva, BARRO, Senén and AMORIM, Dinani, 2014. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research. 2014. Vol. 15, no. 1, p. 3133–3181.

FLATH, Christoph M. and STEIN, Nikolai, 2018. Towards a data science toolbox for industrial analytics applications. Computers in Industry. January 2018. Vol. 94, p. 16–25. DOI 10.1016/j.compind.2017.09.003.

FOUAD, Ahmed M., ATIYA, Amir F., SALEH, Mohamed and BAYOUMI, Abd El-Moniem M., 2014. A simulation-based overbooking approach for hotel revenue management. In: Computer Engineering Conference (ICENCO), 2014 10th International [online]. IEEE. 2014. p. 61–69. [Viewed 14 November 2015]. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7050433

FREISLEBEN, B. and GLEICHMANN, G., 1993. Controlling airline seat allocations with neural networks. In: Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences, 1993. January 1993. p. 635–642 vol.4.

GAMA, João, MEDAS, Pedro, CASTILLO, Gladys and RODRIGUES, Pedro, 2004. Learning with drift detection. In: In SBIA Brazilian Symposium on Artificial Intelligence. Springer Verlag. 2004. p. 286–295.

GARROW, Laurie and FERGUSON, Mark, 2008. Revenue management and the analytics explosion: Perspectives from industry experts. Journal of Revenue and Pricing Management. June 2008. Vol. 7, no. 2, p. 219–229. DOI 10.1057/rpm.2008.3.

GAYAR, Neamat Farouk El, SALEH, Mohamed, ATIYA, Amir, EL-SHISHINY, Hisham, ZAKHARY, Athanasius Alkes Youhanna Fayez and HABIB, Heba Abdel Aziz Mohammed, 2011. An integrated framework for advanced hotel revenue management. International Journal of Contemporary Hospitality Management. 2011. Vol. 23, no. 1, p. 84–98. DOI 10.1108/09596111111101689.

GREENHALGH, Trisha and PEACOCK, Richard, 2005. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ. 5 November 2005. Vol. 331, no. 7524, p. 1064–1065. DOI 10.1136/bmj.38636.593461.68.

GRUN, Bettina and HORNIK, Kurt, 2011. topicmodels: An R Package for fitting topic Models. Journal of Statistical Software. 2011. Vol. 40, no. 11, p. 1–30. DOI 10.18637/jss.v040.i13.

GUERREIRO, João, RITA, Paulo and TRIGUEIROS, Duarte, 2016. A text mining-based review of cause-related marketing literature. Journal of Business Ethics. November 2016. Vol. 139, no. 1, p. 111–128. DOI 10.1007/s10551-015-2622-4.

GÜNTHER, Wendy Arianne, REZAZADE MEHRIZI, Mohammad H., HUYSMAN, Marleen and FELDBERG, Frans, 2017. Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems. September 2017. Vol. 26, no. 3, p. 191–209. DOI 10.1016/j.jsis.2017.07.003.

GUO, Xiaolong, DONG, Yufeng and LING, Liuyi, 2016. Customer perspective on overbooking: The failure of customers to enjoy their reserved services, accidental or intended? Journal of Air Transport Management. June 2016. Vol. 53, p. 65–72. DOI 10.1016/j.jairtraman.2016.01.001.

GUYON, Isabelle and ELISSEEFF, André, 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003. Vol. 3, p. 1157–1182.

HALKOS, George E. and TSILIKA, Kyriaki D., 2015. A framework for stochastic simulation of distribution practices for hotel reservations. In: [online]. 2015. p. 850117. [Viewed 25 July 2018]. Available from: http://aip.scitation.org/doi/abs/10.1063/1.4913172

HANEEM, Faizura, KAMA, Nazri, ALI, Rosmah and SELAMAT, Ali, 2017. Applying data analytics approach in systematic literature review: Master data management case study. In: Frontiers in Artificial Intelligence and Applications. Kitakyushu, Japan. 2017. p. 705–715.

HASTIE, Trevor, TIBSHIRANI, Robert and FRIEDMAN, Jerome, 2001. The elements of statistical learning [online]. Springer series in statistics Springer, Berlin. [Viewed 7 June 2015]. Available from: http://statweb.stanford.edu/~tibs/book/preface.ps

HAYES, David K. and MILLER, Allisha A., 2011. Revenue management for the hospitality industry. Hoboken, NJ, USA: John Wiley & Sons, Inc. ISBN 978-0-470-39308-6.

HEVNER, Alan R., MARCH, Salvatore T., PARK, Jinsoo and RAM, Sudha, 2004. Design science in information systems research. MIS Quarterly. March 2004. Vol. 28, no. 1, p. 75–105.

HORNIK, Kurt, 2017. NLP: Natural language processing Infrastructure [online]. Available from: https://CRAN.R-project.org/package=NLP

HOTREC - ASSOCIATION OF HOTELS, RESTAURANTS AND CAFES AND SIMILAR ESTABLISHMENTS OF EUROPE, 2016. Dominant online platforms gaining market share in travel trade, no signs of increased competition between online travel agents – unveils European hotel distribution study. hotrec.eu [online]. 18 July 2016. [Viewed 24 August 2016]. Available from: https://www.hotrec.eu/wp-content/customer-area/storage/b47b7e97129e1b27c18d8968cb252f5f/Dominant-online-platforms-gaining-market-share-in-travel-trade-no-signs-of-increased-competition-between-online-travel-agents-unveils-European-hotel-distribution-study-18-july-2016.pdf

HUANG, Han-Chen, CHANG, Allen Y. and HO, Chih-Chung, 2013. Using artificial neural networks to establish a customer-cancellation prediction model. Przeglad Elektrotechniczny. 2013. Vol. 89, no. 1b, p. 178–180.

HUEGLIN, Christoph and VANNOTTI, Francesco, 2001. Data mining techniques to improve forecast accuracy in airline business. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining [online]. San Francisco, CA, USA: ACM. 2001. p. 438–442. [Viewed 14 October 2015]. Available from: http://dl.acm.org/citation.cfm?id=502578

ILIESCU, Dan C., GARROW, Laurie A. and PARKER, Roger A., 2008. A hazard model of US airline passengers' refund and exchange behavior. Transportation Research Part B: Methodological. March 2008. Vol. 42, no. 3, p. 229–242.

ILIESCU, Dan C., 2008. Customer based time-to-event models for cancellation behavior: A revenue management integrated approach. Georgia Tech, USA.

IMPERVA, 2014. Detecting and Blocking Site Scraping Attacks. Protecting Valuable Data and Intellectual Property from Online Thieves [online]. 2014. [Viewed 13 December 2015]. Available from: http://www.imperva.com/docs/wp_detecting_and_blocking_site_scraping_attacks.pdf

INSTITUTO NACIONAL DE ESTATÍSTICA, 2016. Tourism statistics-2015. Estatísticas do Turismo [online]. July 2016. [Viewed 1 September 2016]. Available from: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=265858123&PUBLICACOEStema=55581&PUBLICACOESmodo=2

INTERNATIONAL CIVIL AVIATION ORGANIZATION, 2010. Guidelines on Passenger Name Record (PNR) data. [online]. 2010. [Viewed 17 February 2016]. Available from: https://www.iata.org/iata/passenger-data-toolkit/assets/doc_library/04-pnr/New%20Doc%209944%201st%20Edition%20PNR.pdf

IVANOV, Stanislav, 2014. Hotel revenue management: From theory to practice. Varna, Bulgary: Zangador.

IVANOV, Stanislav and ZHECHEV, Vladimir, 2012. Hotel revenue management–A critical literature review. Turizam: znanstveno-strucnicasopis. 2012. Vol. 60, no. 2, p. 175–197.

JENNINGS, Frank and YATES, John, 2009. Scrapping over data: are the data scrapers' days numbered? Journal of Intellectual Property Law & Practice. 2009. Vol. 4, no. 2, p. 120–129. DOI 0.1093/jiplp/jpn232.

JONES, Peter and CHEN, Meng-Mei, 2011. Factors determining hotel selection: Online behaviour by leisure travellers. Tourism and Hospitality Research. 1 January 2011. Vol. 11, no. 1, p. 83–95. DOI 10.1057/thr.2010.20.

KAHN, Matthew E. and LIU, Peter, 2016. Utilizing "Big Data" to Improve the Hotel Sector's Energy Efficiency: Lessons from Recent Economics Research. Cornell Hospitality Quarterly. 2016. Vol. 57, no. 2, p. 202–210.

KASSAMBARA, Alboukadel and MUNDT, Fabian, 2017. factoextra: Extract and visualize the results of multivariate data analyses [online]. Available from: https://CRAN.R-project.org/package=factoextra

KASSAMBARA, Alboukadel, 2017. Practical guide to cluster analysis in R: Unsupervised machine learning. STHDA.

KIMES, Sheryl E. and WIRTZ, Jochen, 2003. Has revenue management become acceptable? Findings from an International study on the perceived fairness of rate fences. Journal of Service Research. 11 January 2003. Vol. 6, no. 2, p. 125–135. DOI 10.1177/1094670503257038.

KIMES, Sheryl E., 2010. The future of hotel revenue management. Cornell Hospitality Reports [online]. 2010. Vol. 10, no. 14. [Viewed 27 October 2013]. Available from:

https://www.hotelschool.cornell.edu/chr/pdf/showpdf/1535/chr/research/kimesrmfuture.pdf

KOHAVI, Ron and LONGBOTHAM, Roger, 2017. Online controlled experiments and A/B tests. Encyclopedia of machine learning and data mining [online]. New York, NY: Springer Science+Business Media. [Viewed 14 August 2017]. Available from: http://www.academia.edu/download/40938549/2015_Online_Controlled_Experiments_EncyclopediaOfMLDM.pdf

KRASTEVA, Rouska, 2017. Local impact of refugee and migrants crisis on greek tourism industry. Economic Studies journal. 2017. No. 4, p. 182–195.

KUHN, Max and JOHNSON, Kjell, 2013. Applied predictive modeling. New York, NY: Springer New York. ISBN 978-1-4614-6848-6.

LAN, Yingjie, BALL, Michael O. and KARAESMEN, Itir Z., 2011. Regret in overbooking and fare-class allocation for single leg. Manufacturing & Service Operations Management. April 2011. Vol. 13, no. 2, p. 194–208. DOI 10.1287/msom.1100.0316.

LANYRD.COM, [no date]. Lanyrd - discover thousands of conferences and professional events! [online]. [Viewed 10 February 2018]. Available from: http://lanyrd.com/

LAW, Rob, 2000. Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. Tourism Management. 1 August 2000. Vol. 21, no. 4, p. 331–340. DOI 10.1016/S0261-5177(99)00067-9.

LAWRENCE, Richard D., HONG, Se June and CHERRIER, Jacques, 2003. Passenger-based predictive modeling of airline no-show rates. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining [online]. ACM. 2003. p. 397–406. [Viewed 14 October 2015]. Available from: http://dl.acm.org/citation.cfm?id=956796

LAWRENCE, Richard D., 2003. A machine-learning approach to optimal bid pricing. In: Computational modeling and problem solving in the networked world. Springer US. p. 97–118. Operations Research/Computer Science Interfaces Series, 21. ISBN 978-1-4613-5366-9.

LEE, Anthony Owen, 1990. Airline reservations forecasting: Probabilistic and statistical models of the booking process [online]. Cambridge, Mass.: Flight Transportation Laboratory, Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology,[1990]. [Viewed 28 July 2015]. Available from: http://dspace.mit.edu/handle/1721.1/68100

LEE, Misuk, 2018. Modeling and forecasting hotel room demand based on advance booking information. Tourism Management. 1 June 2018. Vol. 66, p. 62–71. DOI 10.1016/j.tourman.2017.11.004.

LEMKE, Christiane, RIEDEL, Silvia and GABRYS, Bogdan, 2009. Dynamic combination of forecasts generated by diversification procedures applied to forecasting of airline cancellations. In: IEEE Symposium on Computational Intelligence for Financial Engineering, 2009. CIFEr '09. March 2009. p. 85–91.

LEMKE, Christiane, RIEDEL, Silvia and GABRYS, Bogdan, 2013. Evolving forecast combination structures for airline revenue management. Journal of Revenue and Pricing Management. 1 May 2013. Vol. 12, no. 3, p. 221–234. DOI 10.1057/rpm.2012.30.

LEUNG, Daniel, LAW, Rob, HOOF, Hubert van and BUHALIS, Dimitrios, 2013. Social media in tourism and hospitality: A literature review. Journal of Travel & Tourism Marketing. 1 January 2013. Vol. 30, no. 1–2, p. 3–22. DOI 10.1080/10548408.2013.750919.

LEWIS-BECK, Michael S., 2005. Election forecasting: Principles and practice. The British Journal of Politics & International Relations. 2005. Vol. 7, no. 2, p. 145–164.

LIU, Patrick H., 2004. Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In: Revenue management and pricing: Case studies and applications. Cengage Learning EMEA. p. 91–108. ISBN 1-84480-062-8.

LIU, Y., TEICHERT, T., ROSSI, M., LI, H. and HU, F., 2017. Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. Tourism Management. 2017. Vol. 59, p. 554–563. DOI 10.1016/j.tourman.2016.08.012.

MARTIN-FUENTES, Eva and MELLINAS, Juan Pedro, 2018. Hotels that most rely on Booking.com – online travel agencies (OTAs) and hotel distribution channels. Tourism Review [online]. 13 July 2018. [Viewed 28 August 2018]. DOI 10.1108/TR-12-2017-0201. Available from: https://www.emeraldinsight.com/doi/10.1108/TR-12-2017-0201

MATSUO, Yutaka, 2003. Prediction, forecasting, and chance Discovery. In: Chance discovery. Berlin, Heidelberg: Springer. Advance information processing. ISBN 978-3-642-05609-3.

MCGUIRE, Kelly Ann, 2016. Hotel pricing in a social world: driving value in the digital economy. Hoboken, New Jersey: John Wiley & Sons, Inc. The Wiley & SAS business series. ISBN 978-1-119-12996-7.

MCGUIRE, Kelly Ann, 2017. The analytic hospitality executive: implementing data analytics in hotels and casinos. Hoboken, New Jersey: John Wiley & Sons, Inc. Wiley and SAS business series. ISBN 978-1-119-12998-1.

MCNAMARA, Amelia, RUBIA, Eduardo Arino de la, ZHU, Hao, ELLIS, Shannon and QUINN, Michael, 2018. skimr: Compact and flexible summaries of data. R package version 1.0.1 [online]. Available from: https://CRAN.R-project.org/package=skimr

MEHROTRA, Ravi and RUTTLEY, James, 2006. Revenue management (second ed.). Washington, DC, USA: American Hotel & Lodging Association (AHLA).

MELLINAS, Juan Pedro, MARÍA-DOLORES, Soledad-María Martínez and GARCÍA, Juan Jesús Bernal, 2016. Effects of the Booking.com scoring system. Tourism Management. December 2016. Vol. 57, p. 80–83. DOI 10.1016/j.tourman.2016.05.015.

MONKMAN, Graham George, KAISER, Michel and HYDER, Kieran, 2018. The ethics of using social media in fisheries research. Reviews in Fisheries Science & Aquaculture. 3 April 2018. Vol. 26, no. 2, p. 235–242. DOI 10.1080/23308249.2017.1389854.

MORALES, Dolores R. and WANG, Jingbo, 2010. Forecasting cancellation rates for services booking revenue management using data mining. European Journal of Operational Research. 16 April 2010. Vol. 202, no. 2, p. 554–562. DOI 10.1016/j.ejor.2009.06.006.

MORO, Sérgio, CORTEZ, Paulo and RITA, Paulo, 2015. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications. February 2015. Vol. 42, no. 3, p. 1314–1324.

DOI 10.1016/j.eswa.2014.09.024.

MOUNT, John and ZUMEL, Nina, 2017. vtreat: A statistically sound “data.frame” processor/conditioner [online]. [Viewed 30 June 2017]. Available from: https://cran.r-project.org/web/packages/vtreat/index.html

NEULING, Rainer, RIEDEL, Silvia and KALKA, Kai-Uwe, 2004. New approaches to origin and destination and no-show forecasting: Excavating the passenger name records treasure. Journal of Revenue and Pricing management. 1 April 2004. Vol. 3, no. 1, p. 62–72.

NIKITA, Murzintcev, 2016. ldatunning: Tuning of the Latent Dirichlet Allocation model parameters [online]. [Viewed 27 October 2017]. Available from: https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf

NOONE, Breffni M. and LEE, Chung Hun, 2011. Hotel overbooking: The effect of overcompensation on customers’ reactions to denied service. Journal of Hospitality & Tourism Research. 1 August 2011. Vol. 35, no. 3, p. 334–357. DOI https://doi.org/10.1177/1096348010382238.

NUNEZ-MIR, Gabriela C., IANNONE, Basil V., PIJANOWSKI, Bryan C., KONG, Ningning and FEI, Songlin, 2016. Automated content analysis: addressing the big literature challenge in ecology and evolution. Methods in Ecology and Evolution. November 2016. Vol. 7, no. 11, p. 1262–1272. DOI 10.1111/2041-210X.12602.

OANCEA, Octavian, 2014. Pitfalls of airline revenue management observations. Journal of Revenue and Pricing Management. 1 August 2014. Vol. 13, no. 4, p. 334–338. DOI 10.1057/rpm.2014.9.

O’NEIL, Cathy and SCHUTT, Rachel, 2013. Doing data science. Sebastopol, CA, USA: O’Reilly Media. ISBN 1-4493-5865-9.

PADHI, Sidhartha S. and AGGARWAL, Vijay, 2011. Competitive revenue management for fixing quota and price of hotel commodities under uncertainty. International Journal of Hospitality Management. September 2011. Vol. 30, no. 3, p. 725–734. DOI 10.1016/j.ijhm.2010.12.007.

PAN, Bing and YANG, Yang, 2017a. Monitoring and forecasting tourist activities with big data. In: Management science in hospitality and tourism: Theory, practice, and applications [online]. Apple Academic Press. p. 43–62. [Viewed 30 December 2016]. ISBN 978-1-4822-2347-7. Available from: http://www.crcnetbase.com/doi/pdfplus/10.1201/b19937-1

PAN, Bing and YANG, Yang, 2017b. Forecasting destination weekly hotel occupancy with big data. Journal of Travel Research. 2017. Vol. 56, no. 7, p. 957–970.

PARK, Jeong-Yeol and JANG, SooCheong (Shawn), 2014. Sunk costs and travel cancellation: Focusing on temporal cost. Tourism Management. February 2014. Vol. 40, p. 425–435. DOI 10.1016/j.tourman.2013.08.005.

PARK, June Young and NAGY, Zoltan, 2018. Comprehensive analysis of the relationship between thermal comfort and building control research - A data-driven literature review. Renewable and Sustainable Energy Reviews. 1 February 2018. Vol. 82, p. 2664–2679. DOI 10.1016/j.rser.2017.09.102.

PEREIRA, Luis Nobre, 2016. An introduction to helpful forecasting methods for hotel revenue management. International Journal of Hospitality Management. September 2016. Vol. 58,

p. 13–23. DOI 10.1016/j.ijhm.2016.07.003.

PETRARU, Oren, 2016. Airline passenger cancellations: modeling, forecasting and impacts on revenue management [online]. MsC Thesis. Boston, MA, USA: Massachusetts Institute of Technology. [Viewed 24 July 2018]. Available from: http://hdl.handle.net/1721.1/104325

PHILLIPS, Robert L., 2005. Pricing and revenue optimization. Stanford, CA, USA: Stanford University Press. ISBN 978-0-8047-4698-4.

PIATETSKY, Gregory, 2014. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. kdnuggets.com [online]. October 2014. [Viewed 13 August 2018]. Available from: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

PULUGURTHA, Srinivas S. and NAMBISAN, Shashi S., 2003. A decision-support tool for airline yield management using genetic algorithms. Computer-Aided Civil and Infrastructure Engineering. May 2003. Vol. 18, no. 3, p. 214–223. DOI 10.1111/1467-8667.00311.

QUIÑONERO-CANDELA, Joaquin, SUGIYAMA, Masashi, SCHWAIGHOFER, Anton and LAWRENCE, Neil D. (eds.), 2009. Dataset shift in machine learning. Cambridge, Massachusetts: MIT Press. Neural information processing series. ISBN 978-0-262-17005-5.

R CORE TEAM, 2016. R: A language and environment for statistical computing [online]. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/

RABIANSKI, Joseph S., 2003. Primary and secondary data: Concepts, concerns, errors, and issues. Appraisal Journal. January 2003. Vol. 71, no. 1, p. 43 (13).

REITERMANOVÁ, Z, 2010. Data splitting. In: WDS'10 Proceeding of Contributing Papers. Praha: Matfyzpress. 2010. p. 31–36. ISBN 978-80-7378-139-2.

SHMUELI, Galit and KOPPIUS, Otto R., 2011. Predictive analytics in information systems research. Mis Quarterly. 2011. Vol. 35, no. 3, p. 553–572. DOI 10.2307/23042796.

SHMUELI, Galit, 2010. To Explain or to predict? Statistical Science. August 2010. Vol. 25, no. 3, p. 289–310. DOI 10.1214/10-STS330.

SIERAG, D.D., KOOLE, G.M., VAN DER MEI, R.D., VAN DER REST, J.I. and ZWART, B., 2015. Revenue management under customer choice behaviour with cancellations and overbooking. European Journal of Operational Research. October 2015. Vol. 246, no. 1, p. 170–185. DOI 10.1016/j.ejor.2015.04.014.

SMITH, Scott J., PARSA, H.G., BUJISIC, Milos and VAN DER REST, Jean-Pierre, 2015. Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry. Journal of Travel & Tourism Marketing. 3 October 2015. Vol. 32, no. 7, p. 886–906. DOI 10.1080/10548408.2015.1063864.

SMOLA, Alex and VISHWANATHAN, S. V. N., 2008. Introduction to machine learning. Cambridge, UK: Cambridge University Press. ISBN 0-521-82583-0.

SONG, Haiyan and LIU, Han, 2017. Predicting tourist demand using big data. In: Analytics in Smart Tourism Design [online]. Cham: Springer International Publishing. p. 13–29.

[Viewed 14 August 2017]. ISBN 978-3-319-44262-4. Available from: http://link.springer.com/10.1007/978-3-319-44263-1_2

STEKHOVEN, Daniel J., 2013. missForest: Nonparametric missing value imputation using Random Forest [online]. [Viewed 13 February 2018]. Available from: https://cran.r-project.org/web/packages/missForest/index.html

TALLURI, Kalyan T., VAN RYZIN, Garrett J., KARAESMEN, Itir Z. and VULCANO, Gustavo J., 2008. Revenue management: models and methods. In: Proceedings of the 40th Conference on Winter Simulation [online]. Winter Simulation Conference. 2008. p. 145–156. [Viewed 7 June 2015]. Available from: http://dl.acm.org/citation.cfm?id=1516778

TALLURI, Kalyan T. and VAN RYZIN, Garrett, 2005. The theory and practice of revenue management. New York, NY: Springer. International series in operations research & management science, 68. ISBN 1-4020-7701-7.

TENNEKES, Martjn and DE JONGE, Edwin, 2017. tabplot: Tableplot, a visualization of large datasets [online]. Available from: https://CRAN.R-project.org/package=tabplot

TIMEANDDATE.COM, [no date]. About Time and Date AS. [online]. [Viewed 10 February 2018]. Available from: https://www.timeanddate.com/company/

TSAFNAT, Guy, GLASZIOU, Paul, CHOONG, Miew Keen, DUNN, Adam, GALGANI, Filippo and COIERA, Enrico, 2014. Systematic review automation technologies. Systematic Reviews. 9 July 2014. Vol. 3, p. 74. DOI 10.1186/2046-4053-3-74.

TSAI, Tsung-Hsien, 2011. A temporal case-based procedure for cancellation forecasting: a case study. Current Politics and Economics of South, Southeastern, and Central Asia. 2011. Vol. 20, no. 2, p. 159–182.

TSE, Tony S. M. and POON, Yiu-Tung, 2017. Modeling no-shows, cancellations, overbooking, and walk-ins in restaurant revenue management. Journal of Foodservice Business Research. 15 March 2017. Vol. 20, no. 2, p. 127–145. DOI 10.1080/15378020.2016.1198626.

VAN LEEUWEN, Rik, 2018. Cancellation predictor for revenue management - applied in the hospitality industry [online]. 13 February 2018. Vrije Universiteit Amsterdam. [Viewed 24 July 2018]. Available from: https://beta.vu.nl/nl/Images/werkstuk-leeuwen_rik_van_tcm235-876479.pdf

VIGLIA, Giampaolo, MINAZZI, Roberta and BUHALIS, Dimitrios, 2016. The influence of e-word-of-mouth on hotel occupancy rate. International Journal of Contemporary Hospitality Management. 12 September 2016. Vol. 28, no. 9, p. 2035–2051. DOI 10.1108/IJCHM-05-2015-0238.

WALL STREET JOURNAL, [no date]. Stock market data & financial markets summary - Wall Street Journal. Market Data: The Wall Street Journal [online]. [Viewed 23 August 2018]. Available from: https://markets.wsj.com/usoverview

WANG, Jiana-Fu, 2015. Customer lifetime value incorporating negative behavior: Measurement and comparison. In: Proceedings of 4th International Conference on Logistics, Informatics and Service Science [online]. Berlin, Heidelberg, Germany: Springer. 2015. [Viewed 25 July 2018]. ISBN 978-3-662-43870-1. Available from: https://link.springer.com/chapter/10.1007/978-3-662-43871-8_129

WANG, Xuan Lorna, YOONJOUNG HEO, Cindy, SCHWARTZ, Zvi, LEGOHÉREL, Patrick and SPECKLIN, Frédéric, 2015. Revenue management: Progress, challenges, and research prospects. Journal of Travel & Tourism Marketing. 3 October 2015. Vol. 32, no. 7, p. 797–811. DOI 10.1080/10548408.2015.1063798.

WEATHER UNDERGROUND, [no date]. Weather Undergound API. [online]. [Viewed 10 February 2018]. Available from: https://www.wunderground.com/weather/api/

WEATHERFORD, L. R., GENTRY, T. W. and WILAMOWSKI, B., 2003. Neural network forecasting for airlines: A comparative analysis. Journal of Revenue and Pricing Management. 1 January 2003. Vol. 1, no. 4, p. 319–331. DOI 10.1057/palgrave.rpm.5170036.

WEATHERFORD, Larry R. and KIMES, Sheryl E., 2003. A comparison of forecasting methods for hotel revenue management. International Journal of Forecasting. July 2003. Vol. 19, no. 3, p. 401–415. DOI 10.1016/S0169-2070(02)00011-0.

WEATHERFORD, Larry, 2016. The history of forecasting models in revenue management. Journal of Revenue and Pricing Management. July 2016. Vol. 15, no. 3–4, p. 212–221. DOI 10.1057/rpm.2016.18.

WEATHERFORD, Lawrence R, KIMES, Sheryl E and SCOTT, Darren A, 2001. Forecasting for hotel revenue management: Testing aggregation against disaggregation. Cornell Hotel and Restaurant Administration Quarterly. August 2001. P. 53–64.

WEBB, Geoffrey I., HYDE, Roy, CAO, Hong, NGUYEN, Hai Long and PETITJEAN, Francois, 2016. Characterizing concept drift. Data Mining and Knowledge Discovery. July 2016. Vol. 30, no. 4, p. 964–994. DOI 10.1007/s10618-015-0448-4.

WEBSTER, Jane and WATSON, Richard T., 2002. Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly. June 2002. Vol. 26, no. 3, p. xiii–xxiii.

WELBERS, Kasper, VAN ATTEVELDT, Wouter and BENOIT, Kenneth, 2017. Text analysis in R. Communication Methods and Measures. 2 October 2017. Vol. 11, no. 4, p. 245–265. DOI 10.1080/19312458.2017.1387238.

WIRTZ, Jochen, KIMES, Sheryl E., THENG, J. H. and PATTERSON, Paul, 2003. Yield management: Resolving potential customer conflicts. Journal of Revenue and Pricing Management. 2003. Vol. 2, no. 3, p. 216–226.

XIE, Jinhong and GERSTNER, Eitan, 2007. Service escape: Profiting from customer cancellations. Marketing Science. 1 February 2007. Vol. 26, no. 1, p. 18–30.

YANGYONG, ZHU and YUN, Xiong, 2011. Dataology and data science: Up to now. Sciencepaper Online [online]. 16 June 2011. [Viewed 1 January 2014]. Available from: http://www.paper.edu.cn/en_releasepaper/content/4432156

YEOMAN, Ian, 2016. The history of revenue and pricing management – 15 years and more. Journal of Revenue and Pricing Management. 1 July 2016. Vol. 15, no. 3–4, p. 185–196. DOI 10.1057/rpm.2016.36.

ZAKHARY, Athanasius, ATIYA, Amir F., EL-SHISHINY, Hisham and GAYAR, Neamat, 2011. Forecasting hotel arrivals and occupancy using Monte Carlo simulation. Journal of Revenue and Pricing Management. 2011. Vol. 10, no. 4. DOI 10.1057/rpm.2009.42.

ZAKHARY, Athanasius, GAYAR, Neamat El and AHMED, Sanaa El-Ola H., 2010. Exploiting neural networks to enhance trend forecasting for hotels reservations. In: Artificial Neural Networks in Pattern Recognition [online]. Springer, Berlin, Heidelberg. 11 April 2010. p. 241–251. [Viewed 23 July 2018]. Lecture Notes in Computer Science. ISBN 978-3-642-12158-6. Available from: https://link.springer.com/chapter/10.1007/978-3-642-12159-3_22

ZENKERT, David, 2017. No-show forecast using passenger booking data. Lund, Sweden: Lund University.

ZHANG, Y., SHU, S., JI, Z. and WANG, Y., 2015. A Study of the commercial application of big data of the international hotel group in China: Based on the case study of Marriott international. In: 2015 IEEE First International Conference on Big Data Computing Service and Applications. March 2015. p. 412–417.

ZHU, Wen, ZENG, Nancy and WANG, Ning, 2010. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. In: Proceedings of the NESUG Health Care and Life Sciences [online]. Baltimore, MD. 2010. p. 1–9. [Viewed 20 February 2016]. Available from: http://www.floppybunny.org/robin/web/virtualclassroom/dss/articles/sensitivity_specificity_accuracy_CI.pdf

APPENDICES

APPENDIX A – MACHINE LEARNING METRICS

Several metrics are employed in machine learning classification problems for the assessment of models' performance. The list bellow serves as a brief introduction to those metrics and how they are calculated:

Accuracy (Acc.): Measure of outcome correctness. Measures the proportion of true results among the total number of predictions. The formula is as follows: $Acc. = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN}$.

Area Under the Curve (AUC): Measure of success calculated from the area under the plot of true positive rate (TPR) against false positive rate (FPR).

False Negative (FN): The outcome prediction was negative, but the actual value was positive (e.g., the booking was predicted as likely not to cancel, but it was canceled).

False Positive (FP): The outcome prediction was positive, but the actual value was negative (e.g., the booking was predicted as likely to cancel, but it was not canceled).

False Positive Rate (FPR or Fall-out): Measures the probability of a positive prediction result and the actual value being negative (e.g., probability of a booking being predicted as likely to cancel and effectively did not cancel). The formula is as follows: $FPR = \frac{\sum FP}{\sum FP + \sum TN}$.

Precision (Pre.): Measures the proportion of correct positive predictions. The formula is as follows: $Pre. = \frac{\sum TP}{\sum TP + \sum FP}$.

True Negative Rate (TNR or Specificity): Measures the probability of a negative prediction result and the actual value being negative (e.g., probability of a booking being identified as not likely to cancel and effectively did not cancel). The formula is as follows: $TNR = \frac{\sum TN}{\sum TN + \sum FP}$.

True Negative (TN): The outcome prediction was negative, and so was the actual value (e.g., the booking was predicted as likely not to cancel and has been effectively not canceled).

True Positive (TP): The outcome prediction was positive, and so was the actual value (e.g., the booking was predicted as likely to cancel and has been effectively canceled).

True Positive Rate (TPR, Recall or Sensitivity): Measures the probability of a positive prediction result and the actual value being positive (e.g., probability of a booking being identified as likely to cancel and effectively cancel). The formula is as follows: $TPR = \frac{\sum TP}{\sum TP + \sum FN}$.

APPENDIX B - EXPLORATORY MODELS: DATASETS SUMMARY STATISTICS

This appendix presents summary statistics for the four hotels PMS' datasets employed in the exploratory models. Besides the total number of observations and the total number of variables per hotel dataset, there are three tables that summarize the statistics by variable format: factor (categorical), integer and numeric. The tables are composed by the following columns:

- Categorical variables:
  - Variable: variable name;
  - Missing: number of observations with missing or not available values (NA);
  - Complete: number of observations with values;
  - n: total number of observations;
  - n unique: number of distinct levels/categories;
  - Top counts: top counts by levels/categories;
  - Ordered: indication if the categories are a rank (TRUE) or not (FALSE).
- Integer and numeric variables:
  - Variable: variable name;
  - Missing: number of observations with missing or not available values (NA);
  - Complete: number of observations with values;
  - n: total number of observations;
  - Mean: mean value of the variable;
  - SD: standard deviation;
  - p0: lower value observed;
  - p25: value observed at percentile 25%, i.e. value bellow which 25% observations may be found;
  - p50: value observed at percentile 50% (also known as median);
  - p75: value observed at percentile 75%;
  - p100: upper value observed;
  - Histogram: graphical representation of the variable distribution.

H1 PMS dataset

Observations: 20522

Variables: 38

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	20522	20522	169	6: 5059, 183: 3519, 29: 1283, 184: 944	FALSE
ArrivalDateDayOfWeek	0	20522	20522	7	Sat: 3606, Mon: 3383, Fri: 3319, Thu: 2893	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
ArrivalDateMonth	0	20522	20522	11	Aug: 3337, Sep: 3159, May: 3016, Jul: 2861	FALSE
AssignedRoomType	0	20522	20522	7	A: 9510, B: 4168, C: 3830, D: 1773	FALSE
BookingDateDayOfWeek	0	20522	20522	7	Wed: 3722, Tue: 3438, Mon: 3212, Thu: 3131	FALSE
Company	0	20522	20522	74	NUL: 20003, 181: 118, 160: 66, 239: 54	FALSE
Country	0	20522	20522	89	PRT: 7646, GBR: 3013, ESP: 2990, IRL: 1400	FALSE
CustomerType	0	20522	20522	4	Tra: 17050, Tra: 3038, Con: 304, Gro: 130	FALSE
DepositType	0	20522	20522	3	No : 19362, Non: 684, Ref: 476, NA: 0	FALSE
DistributionChannel	0	20522	20522	8	Onl: 6103, Dir: 5305, Off: 3897, Who: 1840	FALSE
IsCanceled	0	20522	20522	2	0: 16941, 1: 3581, NA: 0	FALSE
IsRepeatedGuest	0	20522	20522	2	0: 19066, 1: 1456, NA: 0	FALSE
IsVIP	0	20522	20522	2	0: 20520, 1: 2, NA: 0	FALSE
MarketSegment	0	20522	20522	9	Fam: 17893, Bus: 805, Oth: 783, Spo: 777	FALSE
Meal	0	20522	20522	4	BB: 17607, HB: 2572, SC: 220, FB: 123	FALSE
ReservedRoomType	0	20522	20522	7	A: 10541, B: 4030, C: 3480, D: 1510	FALSE
WasInWaitingList	0	20522	20522	2	0: 20514, 1: 8, NA: 0	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	20522	20522	2.02	1.25	0	2	2	2	99	▇▁▁▁▁▁▁▁
AgeAtBookingDate	0	20522	20522	36.3	21.88	-4	26	39	52	219	▂▇▅▁▁▁▁▁
ArrivalDateDayOfMonth	0	20522	20522	15.82	8.87	1	8	16	24	31	▇▇▇▇▆▇▇▆
ArrivalDateWeekNumber	0	20522	20522	29.49	8.7	1	22	30	37	53	▁▁▆▇▆▇▂▁
ArrivalDateYear	0	20522	20522	2014.09	0.81	2013	2013	2014	2015	2015	▇▁▁▇▁▁▁▇
Babies	0	20522	20522	0.025	0.16	0	0	0	0	2	▇▁▁▁▁▁▁▁
BookingChanges	0	20522	20522	0.34	0.85	0	0	0	0	46	▇▁▁▁▁▁▁▁
CanceledTime	0	20522	20522	6.17	27.7	-1	-1	-1	-1	406	▇▁▁▁▁▁▁▁
Children	0	20522	20522	0.2	0.54	0	0	0	0	10	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	20522	20522	0.0015	0.2	0	0	0	0	29	▇▁▁▁▁▁▁▁
LeadTime	0	20522	20522	48.47	61.1	0	6	28	67	1002	▇▁▁▁▁▁▁▁
LengthOfStay	0	20522	20522	4.88	4.45	0	2	4	7	370	▇▁▁▁▁▁▁▁
PreviousBookingsNotCanceled	0	20522	20522	0.29	3.58	0	0	0	0	98	▇▁▁▁▁▁▁▁
PreviousCancellations	0	20522	20522	0.015	0.15	0	0	0	0	5	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
PreviousStays	0	20522	20522	0.76	4.9	0	0	0	0	105	▇▁▁▁▁▁▁▁
RequiredCarParkingSpaces	0	20522	20522	0.29	0.46	0	0	0	1	7	▇▃▁▁▁▁▁▁
RoomsQuantity	0	20522	20522	1.03	1.08	0	1	1	1	90	▇▁▁▁▁▁▁▁
StaysInWeekendNights	0	20522	20522	1.32	1.43	0	0	1	2	106	▇▁▁▁▁▁▁▁
StaysInWeekNights	0	20522	20522	3.56	3.19	0	2	3	5	264	▇▁▁▁▁▁▁▁
TotalOfSpecialRequests	0	20522	20522	1.07	1.17	0	0	1	2	5	▇▆▁▂▂▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	20522	20522	91.41	100.64	-3848	51.8	79.4	125	6120	▁▁▁▇▁▁▁▁

H2 PMS dataset

Observations: 9809

Variables: 38

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	9809	9809	73	94: 1274, 3: 1057, 43: 992, 57: 806	FALSE
ArrivalDateDayOfWeek	0	9809	9809	7	Sat: 2055, Sun: 1660, Wed: 1649, Mon: 1426	FALSE
ArrivalDateMonth	0	9809	9809	12	Aug: 1576, Jul: 1262, Sep: 1207, May: 1103	FALSE
AssignedRoomType	0	9809	9809	9	A: 3857, B: 2729, C: 1355, D: 1339	FALSE
BookingDateDayOfWeek	0	9809	9809	7	Mon: 2006, Thu: 1678, Tue: 1616, Fri: 1575	FALSE
Company	0	9809	9809	73	94: 1160, 3: 1013, 43: 970, 57: 792	FALSE
Country	0	9809	9809	38	D: 2523, PRT: 2416, NLD: 1790, GBR: 1042	FALSE
CustomerType	0	9809	9809	3	Con: 6732, Gro: 2828, Tra: 249, NA: 0	FALSE
DepositType	0	9809	9809	3	No : 8459, Ref: 1154, Non: 196, NA: 0	FALSE
DistributionChannel	0	9809	9809	7	Spe: 4804, Dir: 1531, Sta: 1420, Und: 695	FALSE
IsCanceled	0	9809	9809	2	0: 8834, 1: 975, NA: 0	FALSE
IsRepeatedGuest	0	9809	9809	2	0: 7251, 1: 2558, NA: 0	FALSE
IsVIP	0	9809	9809	2	0: 9808, 1: 1, NA: 0	FALSE
MarketSegment	0	9809	9809	2	Und: 9805, Dir: 4, NA: 0	FALSE
Meal	0	9809	9809	3	SC: 7243, BB: 1410, HB: 1156, NA: 0	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
ReservedRoomType	0	9809	9809	9	A: 4003, B: 2098, C: 2049, D: 795	FALSE
WasInWaitingList	0	9809	9809	1	0: 9809, NA: 0	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	9809	9809	2.51	0.94	0	2	2	3	8	▁▇▂▁▁▁▁▁
AgeAtBookingDate	0	9809	9809	40.37	26.81	-58	19	47	61	222	▁▅▅▇▁▁▁▁
ArrivalDateDayOfMonth	0	9809	9809	15.85	9.12	1	8	16	24	31	▇▇▆▇▆▇▇▇
ArrivalDateWeekNumber	0	9809	9809	28.66	11.57	1	20	30	37	53	▁▂▃▇▇▇▂▁
ArrivalDateYear	0	9809	9809	2014.06	0.82	2013	2013	2014	2015	2015	▇▁▁▇▁▁▁▇
Babies	0	9809	9809	0.035	0.19	0	0	0	0	2	▇▁▁▁▁▁▁▁
BookingChanges	0	9809	9809	0.31	0.73	0	0	0	0	10	▇▁▁▁▁▁▁▁
CanceledTime	0	9809	9809	4.47	25.83	-1	-1	-1	-1	447	▇▁▁▁▁▁▁▁
Children	0	9809	9809	0.39	0.76	0	0	0	0	4	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	9809	9809	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
LeadTime	0	9809	9809	109.37	115.1	0	27	79	154	867	▇▃▁▁▁▁▁▁
LengthOfStay	0	9809	9809	10.03	8.81	0	7	7	14	151	▇▁▁▁▁▁▁▁
PreviousBookingsNotCanceled	0	9809	9809	0.55	1.5	0	0	0	0	18	▇▁▁▁▁▁▁▁
PreviousCancellations	0	9809	9809	0.1	0.31	0	0	0	0	3	▇▁▁▁▁▁▁▁
PreviousStays	0	9809	9809	7.83	28.85	0	0	0	0	574	▇▁▁▁▁▁▁▁
RequiredCarParkingSpaces	0	9809	9809	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
RoomsQuantity	0	9809	9809	1	0	1	1	1	1	1	▁▁▁▇▁▁▁▁
StaysInWeekendNights	0	9809	9809	2.85	2.59	0	2	2	4	44	▇▁▁▁▁▁▁▁
StaysInWeekNights	0	9809	9809	7.18	6.28	0	5	5	10	108	▇▁▁▁▁▁▁▁
TotalOfSpecialRequests	0	9809	9809	0.049	0.29	0	0	0	0	4	▇▁▁▁▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	9809	9809	82.4	60.52	-269.28	40	67.44	117	397.31	▁▁▁▇▆▂▁▁

H3 PMS dataset

Observations: 9365

Variables: 38

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	9365	9365	102	177: 1042, 19: 968, 21: 830, 3: 783	FALSE
ArrivalDateDayOfWeek	0	9365	9365	7	Sat: 1937, Sun: 1424, Fri: 1313, Mon: 1287	FALSE
ArrivalDateMonth	0	9365	9365	12	Aug: 1485, Sep: 1126, Jun: 1099, May: 1048	FALSE
AssignedRoomType	0	9365	9365	13	A: 2523, B: 2195, C: 1372, D: 1317	FALSE
BookingDateDayOfWeek	0	9365	9365	7	Wed: 1654, Thu: 1508, Fri: 1474, Mon: 1459	FALSE
Company	0	9365	9365	120	NUL: 7934, 250: 189, 146: 106, 286: 74	FALSE
Country	0	9365	9365	59	GBR: 4831, DEU: 937, PRT: 844, ESP: 701	FALSE
CustomerType	0	9365	9365	3	Tra: 8200, Tra: 666, Gro: 499, NA: 0	FALSE
DepositType	0	9365	9365	3	No : 7875, Non: 856, Ref: 634, NA: 0	FALSE
DistributionChannel	0	9365	9365	6	Tou: 3998, Dir: 1706, Onl: 1509, Und: 1399	FALSE
IsCanceled	0	9365	9365	2	0: 8286, 1: 1079, NA: 0	FALSE
IsRepeatedGuest	0	9365	9365	2	0: 7259, 1: 2106, NA: 0	FALSE
IsVIP	0	9365	9365	1	0: 9365, NA: 0	FALSE
MarketSegment	0	9365	9365	7	Lei: 8043, Gol: 409, Gro: 328, Com: 317	FALSE
Meal	0	9365	9365	6	BB: 5795, SC: 2904, HBH: 496, HBL: 131	FALSE
ReservedRoomType	0	9365	9365	13	A: 2507, B: 2117, C: 1972, D: 1428	FALSE
WasInWaitingList	0	9365	9365	1	0: 9365, NA: 0	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	9365	9365	2.21	0.84	0	2	2	2	7	▁▂▇▁▁▁▁▁
AgeAtBookingDate	0	9365	9365	37.73	23.28	-57	27	43	54	114	▁▁▅▁▇▆▁▁
ArrivalDateDayOfMonth	0	9365	9365	15.29	9.17	1	7	15	23	31	▇▇▇▇▆▇▇▇
ArrivalDateWeekNumber	0	9365	9365	28.92	11.87	1	20	30	37	53	▁▂▃▅▅▆▃▁
ArrivalDateYear	0	9365	9365	2014.07	0.81	2013	2013	2014	2015	2015	▆▁▁▇▁▁▁▇
Babies	0	9365	9365	0.062	0.26	0	0	0	0	3	▇▁▁▁▁▁▁▁
BookingChanges	0	9365	9365	0.43	0.91	0	0	0	1	29	▇▁▁▁▁▁▁▁
CanceledTime	0	9365	9365	3.96	23.89	-1	-1	-1	-1	374	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Children	0	9365	9365	0.36	0.75	0	0	0	0	5	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	9365	9365	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
LeadTime	0	9365	9365	62.36	72.14	0	12	37	86	695	▇▂▁▁▁▁▁▁
LengthOfStay	0	9365	9365	7.61	8.98	0	4	6	8	261	▇▁▁▁▁▁▁▁
PreviousBookingsNotCanceled	0	9365	9365	0.7	2.79	0	0	0	0	56	▇▁▁▁▁▁▁▁
PreviousCancellations	0	9365	9365	0.13	0.37	0	0	0	0	4	▇▁▁▁▁▁▁▁
PreviousStays	0	9365	9365	7.22	26.82	0	0	0	0	431	▇▁▁▁▁▁▁▁
RequiredCarParkingSpaces	0	9365	9365	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
RoomsQuantity	0	9365	9365	1	0.018	0	1	1	1	2	▁▁▁▇▁▁▁▁
StaysInWeekendNights	0	9365	9365	2.15	2.65	0	1	2	2	74	▇▁▁▁▁▁▁▁
StaysInWeekNights	0	9365	9365	5.46	6.41	0	2	5	6	187	▇▁▁▁▁▁▁▁
TotalOfSpecialRequests	0	9365	9365	0.32	0.56	0	0	0	1	3	▇▁▂▁▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	9365	9365	108.47	77.78	0	58	98	151	760	▇▆▂▁▁▁▁▁

H4 PMS dataset

Observations: 33445

Variables: 38

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	33445	33445	231	NUL: 12100, 910: 4019, 240: 3519, 965: 1108	FALSE
ArrivalDateDayOfWeek	0	33445	33445	7	Thu: 5819, Sat: 5328, Mon: 5108, Fri: 4396	FALSE
ArrivalDateMonth	0	33445	33445	12	Aug: 4124, Sep: 3785, Jul: 3749, Oct: 3741	FALSE
AssignedRoomType	0	33445	33445	10	A: 17465, B: 6406, C: 5651, D: 1718	FALSE
BookingDateDayOfWeek	0	33445	33445	7	Thu: 6160, Mon: 5911, Tue: 5856, Fri: 5766	FALSE
Company	0	33445	33445	268	NUL: 30033, 251: 730, 223: 422, 960: 164	FALSE
Country	0	33445	33445	93	PRT: 16096, GBR: 5814, ESP: 3296, IRL: 2392	FALSE
CustomerType	0	33445	33445	4	Tra: 20401, Tra: 8567, Con: 3050, Gro: 1427	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
DepositType	0	33445	33445	3	No : 32729, Non: 711, Ref: 5, NA: 0	FALSE
DistributionChannel	0	33445	33445	4	Und: 20957, TA/: 8526, Dir: 2482, Cor: 1480	FALSE
IsCanceled	0	33445	33445	2	0: 28078, 1: 5367, NA: 0	FALSE
IsRepeatedGuest	0	33445	33445	2	0: 32814, 1: 631, NA: 0	FALSE
IsVIP	0	33445	33445	2	0: 33440, 1: 5, NA: 0	FALSE
MarketSegment	0	33445	33445	7	Off: 10614, Dir: 7611, Onl: 7045, Gro: 5423	FALSE
Meal	0	33445	33445	5	BB: 25154, HB: 6774, FB: 1128, SC: 219	FALSE
ReservedRoomType	0	33445	33445	10	A: 20525, C: 8240, B: 2455, D: 1279	FALSE
WasInWaitingList	0	33445	33445	2	0: 33338, 1: 107, NA: 0	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	33445	33445	1.91	2.26	0	2	2	2	99	▇▁▁▁▁▁▁▁
AgeAtBookingDate	0	33445	33445	38.21	24.18	-30	25	42	57	167	▁▅▇▇▂▁▁▁
ArrivalDateDayOfMonth	0	33445	33445	15.81	8.93	1	8	16	24	31	▇▇▇▇▆▇▇▇
ArrivalDateWeekNumber	0	33445	33445	29.18	12.44	1	20	30	39	53	▂▃▅▇▇▇▅▂
ArrivalDateYear	0	33445	33445	2014.07	0.82	2013	2013	2014	2015	2015	▆▁▁▇▁▁▁▇
Babies	0	33445	33445	0.0071	0.088	0	0	0	0	5	▇▁▁▁▁▁▁▁
BookingChanges	0	33445	33445	0.25	0.77	0	0	0	0	31	▇▁▁▁▁▁▁▁
CanceledTime	0	33445	33445	5.83	28.39	-5	-1	-1	-1	389	▇▁▁▁▁▁▁▁
Children	0	33445	33445	0.06	0.4	0	0	0	0	50	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	33445	33445	0.29	5.39	0	0	0	0	122	▇▁▁▁▁▁▁▁
LeadTime	0	33445	33445	54.15	71.58	0	3	21	84	744	▇▂▁▁▁▁▁▁
LengthOfStay	0	33445	33445	4.54	8.08	0	2	4	7	659	▇▁▁▁▁▁▁▁
PreviousBookingsNotCanceled	0	33445	33445	0.94	3.42	0	0	0	1	104	▇▁▁▁▁▁▁▁
PreviousCancellations	0	33445	33445	0.029	0.25	0	0	0	0	15	▇▁▁▁▁▁▁▁
PreviousStays	0	33445	33445	2.33	13.54	0	0	0	1	1173	▇▁▁▁▁▁▁▁
RequiredCarParkingSpaces	0	33445	33445	0.05	0.22	0	0	0	0	2	▇▁▁▁▁▁▁▁
RoomsQuantity	0	33445	33445	1.1	1.59	0	1	1	1	99	▇▁▁▁▁▁▁▁
StaysInWeekendNights	0	33445	33445	1.22	2.4	0	0	1	2	189	▇▁▁▁▁▁▁▁
StaysInWeekNights	0	33445	33445	3.32	5.78	0	1	3	5	470	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
TotalOfSpecialRequests	0	33445	33445	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	33445	33445	73.09	94.65	0	39.6	58	90.9	5333	▇▁▁▁▁▁▁▁

APPENDIX C – FEATURES DESCRIPTION

The following table presents a list of all the features available in the different datasets used and/or engineered to employ in the different models developed throughout the dissertation's research work. In addition to a description of each feature, the format, data source(s) of origin and models where it was employed are also detailed. Columns by type, data source, and models can assume the following codes:

- Type:
  - Format:
    - C – Categorical;
    - D – Date;
    - N – Numeric/Integer.
  - Creation:
    - E – Engineered – calculated from other input features;
    - I – Input – no transformation;
    - X – Engineered at extraction – calculated from other features at the time of database extraction.
- Data source:
  - H – Holiday calendar;
  - O – Online prices/inventory;
  - P – PMS;
  - R – Social media reputation;
  - S – Special events calendar;
  - W – Weather forecast.
- Models:
  - 1 – 4 – Final models 1 to 4;
  - E – Exploratory models;
  - T – Field test models;
  - () – used to construct modelling datasets or new features, but not used as features of the models.

Feature	Type	Data source	Models	Description
ADR	N, I	P	E	Average daily rate
Adults	N, I	P	E,1-4, T	Number of adults
AgeAtBookingDate	N, X	P	-	Age in years of the booking holder at the time of booking
Agent	C, I	P	E,1-4, T	ID of agency (if booked through an agency)
ArrivalDateDayOfMonth	N, I	P	E	Number of the day of the month of the arrival date
ArrivalDateMonth	C, I	P	E	Name of month of arrival date
ArrivalDateWeekNumber	N, I	P	E	Week number of the arrival date
ArrivalDateYear	N, I	P	E	Year number of the arrival date
AssignedRoomType	C, I	P	-	Room type assigned to booking
AssocitatedToEvent	C, X	P	4	Binary value indicating if the booking was associated with an event held at the hotel (e.g., meeting or wedding) (0: no; 1: yes)
AVGCompSetNormalizedRating	N, X	R	-	Average normalized ratings of the competitors from Booking.com and Tripadvisor.com, calculated with the min-max formula: $x' = \frac{(x - \min(x))}{(\max(x) - \min(x))} \times 100$

Feature	Type	Data source	Models	Description
AVGNormalizedRating	N, X	R	-	Average normalized ratings of the hotel from Booking.com and Tripadvisor.com, calculated with the min-max formula: $x' = \frac{(x - \min(x))}{(\max(x) - \min(x))} \times 100$
AvgQuantityOfPrecipitationInMM	N, E	W, P	3, 4	Average quantity of precipitation forecasted. This value is calculated by summing the milliliters of precipitation forecasted for each days of the stay, and dividing by the number of days of the stay to which there was a weather forecast. The booking cancellation outcome date is defined as the observation date. For not canceled bookings, the arrival date was considered. For canceled bookings, the cancellation date was considered
AvgWindInKph	N, I	W	-	Average wind in Km per hour
Babies	N, I	P	E, 1-4, T	Number of babies
BookingChanges	N, X	P	E, 1-4, T	Heuristic created by summing the number of booking changes (amendments) prior to arrival date that could indicate cancellation intentions (arrival or departure dates, number of persons, type of meal, ADR, or reserved room type)
BookingDateDayOfWeek	C, X	P	-	Day of week of booking date (Monday to Sunday)
BookedSPA	C, X	P	4	Binary value indicating if a SPA service was booked prior to the guest arrival (0: no; 1: yes)
CanceledTime	N, X	P	-	Number of days prior to expected arrival date that the booking was canceled (if not canceled the value was set to -1)
Children	N, I	P	E, 1-4, T	Number of children
Company	C, I	P	E, 1-4, T	ID of company/corporation (if an account was associated with it)
CompSetMaxAvailableRooms	N, X	O	-	Maximum number of rooms available for sale by any of the five competitors for the same or superior type of meal and maximum occupation
Country	C, I	P	E, 1-4	Country ISO identification of the main booking holder
CustomerType	C, X	P	E, 1-4, T	Type of customer (group, contract, transient, or transient-party); this last category is a heuristic built when the booking is transient but is fully or partially paid in conjunction with other bookings (e.g., small groups such as families who require more than one room)
Date	D, I	H, S, R	(3, 4)	Date
DayOfYear	N, E	P	1-4	Number representing the sequential day of the year. For example, the January 1st is 1 and the February 1st is 32.
DaysInWaitingList	N, X	P	E, 1-4, T	Number of days the booking was in a waiting list prior to confirmed availability and to being confirmed as a booking
DepositType	C, X	P	E, 1-4, T	Since hotels had different cancellation and deposit policies, a heuristic was developed to define the deposit type (nonrefundable, refundable, no deposit): payment made in full before the arrival date was considered a “nonrefundable” deposit, partial payment before arrival was considered a “refundable” deposit, otherwise it was considered as “no deposit”
Description	C, I	H	-	Name of holiday
Designation	C, I	S	-	Name of event

Feature	Type	Data source	Models	Description
DistributionChannel	C, I	P	E, 1-4, T	Distribution channel used to make the booking
FolioNumber	N, I	P	(1-4, T)	PMS booking number
HotelCommonID	C, I	R	(3, 4)	Hotel ID
HotelID	C, I	O	(3, 4)	Hotel ID
HotelsWithOpenSales	N, X	O	(3, 4)	Number of competitors which had rooms on sale
HotelsWithRoomsAvailable	N, E	O, P	1-4	Number of competitors that have inventory for sale for the period of the booking stay, with the same type of meal package, and that could accommodate the equal or superior number of adults. Inventory availability is obtained at the arrival or cancellation date, accordingly to the cancellation outcome
IsRepeatedGuest	C, X	P	E, 1-4, T	Binary value indicating if the booking holder, at the time booking creation, was a repeat guest at the hotel (0: no; 1: yes); created by comparing the time of booking with the guest profile creation record
IsVIP	C, I	P	-	Binary value indicating if the guest should be considered a Very Important Person (0: no; 1: yes)
LeadTime	N, X	P	E, 1-4	Number of days prior to arrival that the hotel received the booking
LengthOfStay	N, I	P	-	Total number of nights the customer booked
LiveTime	N, E	P	T	Number of days from booking creation according to the booking status: for “A” type bookings (not canceled), it was calculated as the number of days between booking creation and arrival; for “B” bookings (canceled), the elapsed number of days between the date of booking creation and the cancellation date was employed; for “C” bookings (unknown outcome), the elapsed number of days between the date of creation and the processing date (current date)
Location	C, I	S, W	(3,4)	Location of the event or weather forecast
LookupDate	D, I	O, W	(3, 4)	Date equal or posterior to observation date at which for which prices and inventory availability (or weather forecast) is checked
MarketSegment	C, I	P	E, 1-4, T	Market segment to which the booking was classified as
MaxAvailableRooms	N, X	O	-	Maximum number of rooms the hotel had available for sale for the same or superior type of meal, and maximum occupation
MaxOccupation	N, X	O	(3, 4)	Maximum number of adults that the room can accommodate
MaxTemperatureInCelsius	N, I	W	-	Maximum forecasted temperature in degrees Celsius
Meal	C, I	P	E, 1-4, T	ID of meal the guest requested
Meal	C, X	O	(3, 4)	Meal included in the rate/price
MealNumber	N, X	O	(3, 4)	Meal ID converted to a rank (1: self-catering; 2: bed and breakfast; 3: half-board; 4: full-board;)
MedianCompSetAVGNormalizedRating	N, X	R	-	Median of the competitors normalized average ratings from Booking.com and Tripadvisor.com
MedianCompSetPrice	N, X	O	(3, 4)	Median price of the competitive set for the same day, for the same or superior type of meal, and maximum occupation

Feature	Type	Data source	Models	Description
MinCompSetPrice	N, X	O	(3, 4)	Minimum price of the competitive set for the same day, for the same or superior type of meal, and maximum occupation
MinPrice	N, X	O	(3, 4)	Hotel minimum price for the same day, for the same or superior type of meal, and maximum occupation
nHolidays	N, E	H, P	1-4	Number of local holidays that are due to occur during the booking period of stay (including the check-out date)
ObservationDate	D, I	O, W	(3, 4)	Date when the prices and inventory availability (or weather forecast) for a future date where observed
PreviousBookingsNotCanceled	N, X	P	(E, 1-4, T)	Number of previous bookings to this booking the guest had that were not canceled
PreviousCancellations	N, X	P	(E, 1-4, T)	Number of previous bookings to this booking the guest had that were canceled
PreviousCancellationRatio	N, E	P	E, 1-4, T	Ratio created by the division of the guest's number of previous cancellations by the guest's previous number of bookings at the hotel
PreviousStays	N, X	P	-	Number of nights the guest had stayed at the hotel prior to the current booking
ProbabilityOfPrecipitation	N, I	W	-	Percentage of probability that it would rain
QuantityOfPrecipitationInMM	N, I	W	(3, 4)	Forecasted quantity of rain it will fall on the lookup date (in millimeters)
RateCode	C, I	P	-	Code of the first night rate
RatioADRbyCompsetMedianDifference	N, E	O, P	1-4	Ratio calculated by the division of the booking ADR, by the average of the median of each of the competitor hotels, for the cheapest room price each competitor had available, with the same type of meal package, that could accommodate the number of adults on the booking, for the same period of stay. Competitor prices are obtained at arrival or cancellation date, accordingly to the cancellation outcome
RatioMajorEventsNights	N, E	S, P	1-4	Ratio calculated by the division of the total number of major special events that are supposed to occur during the stay, by the total number of nights of the booking
RatioMinorEventsNights	N, E	S, P	1-4	Ratio calculated by the division of the total number of minor special events that are supposed to occur during the stay, by the total number of nights of the booking
RequiredCarParkingSpaces	N, I	P	E, 4	Number of car parking spaces required by the guest
ReservationStatus	C, I	P	(1-4, T)	Identification code of the status of the booking (A: canceled; C: Confirmed; G: Guarantee; N: No-show; O: Checked-out; R: Checked-in)
ReservationStatusDate	D, I	P	(1-4, T)	Date when the booking was changed to the current status
ReservedRoomType	C, I	P	E, 1-4	Room type requested by the guest
RoomsQuantity	N, I	P	-	Number of rooms booked
SRDoubleBed	C, X	P	4	Binary value indicating if guest, prior to arrival asked specifically for a double bed (0: no; 1: yes)
SRHighFloor	C, X	P	4	Binary value indicating if guest, prior to arrival asked specifically for a a high floor room (0: no; 1: yes)

Feature	Type	Data source	Models	Description
SRQuietRoom	C, X	P	4	Binary value indicating if guest, prior to arrival asked specifically for a quiet room (0: no; 1: yes)
SRTogether	C, X	P	4	Binary value indicating if guest, prior to arrival asked specifically to be in a room closer to other booking room (0: no; 1: yes)
SRTwinBed	C, X	P	4	Binary value indicating if guest, prior to arrival asked specifically for a twin bed (0: no; 1: yes)
StaysInWeekendNights	N, X	P	E, 1-4, T	From the total length of stay, how many nights were in weekends (Saturday and Sunday)
StaysInWeekNights	N, X	P	E, 1-4, T	From the total length of stay, how many nights were in weekdays (Monday through Friday)
SUMTotalReviewsOnSite	N, X	R	-	Total number of reviews published on both Tripadvisor.com and Booking.com
ThirdQuartileDeviationADR	N, E	P	1-4, T	Ratio calculated by the division of the booking ADR by the third quartile value, of all bookings of the same distribution channel, same reserved room type, for the same expected week/year of arrival
TotalOfSpecialRequests	N, X	P	E, 1-4, T	Number of special requests made (e.g. fruit basket, sea view, etc.)
Type	C, X	S	(3, 4)	Category of the event (minor: events with local impact, e.g. conferences; major: events with regional or national impact, e.g. music festivals;)
WasInWaitingList	N, I	P	T	Binary value that indicates if the booking was entered on a waiting list or directly entered as a booking (0: normal booking; 1: waiting list)
WorseThan	N, X	O	-	Number of competitors which had lower prices for the same or superior meal, and maximum occupation
WorseThan	N, X	R, P	3, 4	Number of hotels from the competition set who had a better rating at arrival or cancellation date, accordingly to the booking cancellation outcome

APPENDIX D - FINAL MODELS: DATASETS SUMMARY STATISTICS

This appendix presents the summary statistics of the eight hotels PMS' datasets employed in the final models and also of the additional data sources. Besides the total number of observations and the total number of variables per dataset, four tables summarize statistics by variable type: date, factor (categorical), integer and numeric. These tables are composed by the following columns:

- Date variables:
    - Variable: variable name;
    - Missing: number of observations with missing or not available values (NA);
    - Complete: number of observations with values;
    - n: total number of observations;
    - Min: older date found on an observation;
    - Max: younger date found on an observation;
    - Median: if order by date and divided in two sample of observations, this value would be the date that would separate the older dates sample, from the younger dates sample;
    - n unique: number of distinct dates present.
- Categorical variables:
    - Variable: variable name;
    - Missing: number of observations with missing or not available values (NA);
    - Complete: number of observations with values;
    - n: total number of observations;
    - n unique: number of distinct levels/categories;
    - Top counts: top counts by levels/categories;
    - Ordered: indication if the categories are a rank (TRUE) or not (FALSE).
- Integer and numeric variables:
    - Variable: variable name;
    - Missing: number of observations with missing or not available values (NA);
    - Complete: number of observations with values;
    - n: total number of observations;
    - Mean: mean value of the variable;
    - SD: standard deviation;
    - p0: lower value observed;
    - p25: value observed at percentile 25%, i.e. value bellow which 25% observations may be found;
    - p50: value observed at percentile 50% (also known as median);
    - p75: value observed at percentile 75%;
    - p100: upper value observed;

- Histogram: graphical representation of the variable distribution.

C1 PMS dataset

Observation: 75337

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	75337	75337	2016-01-01	2017-11-28	2016-12-13	698

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	75337	75337	215	9: 33406, NUL: 8195, 14: 3901, 7: 3780	FALSE
ArrivalDateMonth	0	75337	75337	12	Oct: 9655, Sep: 8478, May: 7952, Jun: 7544	FALSE
ArrivalDateMonthYear	0	75337	75337	23	201: 5605, 201: 4870, 201: 4556, 201: 4050	FALSE
AssignedRoomType	0	75337	75337	9	A: 52864, D: 15284, E: 2333, F: 2039	FALSE
Company	0	75337	75337	217	NUL: 71873, 40: 815, 153: 239, 45: 207	FALSE
Country	0	75337	75337	169	PRT: 24593, FRA: 8581, DEU: 6859, GBR: 6050	FALSE
CustomerType	0	75337	75337	4	Tra: 60991, Tra: 13785, Gro: 402, Con: 159	FALSE
DepositType	0	75337	75337	3	No : 65378, Non: 9926, Ref: 33, NA: 0	FALSE
DistributionChannel	0	75337	75337	4	TA/: 64986, Dir: 6944, Cor: 3133, GDS: 274	FALSE
IsCanceled	0	75337	75337	2	0: 44602, 1: 30735, NA: 0	FALSE
IsRepeatedGuest	0	75337	75337	2	0: 73573, 1: 1764, NA: 0	FALSE
MarketSegment	0	75337	75337	7	Onl: 40793, Off: 14239, Gro: 10709, Dir: 6192	FALSE
Meal	0	75337	75337	4	BB: 58715, SC: 11952, HB: 4661, FB: 9	FALSE
ReservationStatus	0	75337	75337	3	O: 44602, A: 29827, N: 908, NA: 0	FALSE
ReservedRoomType	0	75337	75337	8	A: 57257, D: 12976, E: 1839, F: 1819	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	75337	75337	1.85	0.52	0	2	2	2	4	▁▂▁▇▁▁▁▁
ArrivalDateDay OfMonth	0	75337	75337	15.54	8.68	1	8	15	23	31	▇▇▇▇▆▇▇▆
ArrivalDateWeek Number	0	75337	75337	27.34	13.31	1	17	27	39	53	▃▆▆▇▅▇▇▂
ArrivalDateYear	0	75337	75337	2016.53	0.5	2016	2016	2017	2017	2017	▇▁▁▁▁▁▁▇

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
BookingChanges	0	75337	75337	0.22	0.66	0	0	0	0	21	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	75337	75337	3.31	29.04	0	0	0	0	480	▇▁▁▁▁▁▁▁
FolioNumber	0	75337	75337	60627.59	23218.29	5012	41205	60573	79708	106775	▁▅▇▇▇▇▇▅
LeadTime	0	75337	75337	109.87	116.04	-1	23	73	163	823	▇▃▂▁▁▁▁▁
PreviousBookingsNotCanceled	0	75337	75337	0.19	2.19	0	0	0	0	72	▇▁▁▁▁▁▁▁
Previous Cancellations	0	75337	75337	0.038	0.93	0	0	0	0	38	▇▁▁▁▁▁▁▁
RequiredCar ParkingSpaces	0	75337	75337	0.024	0.15	0	0	0	0	3	▇▁▁▁▁▁▁▁
StaysInWeekend Nights	0	75337	75337	0.82	0.88	0	0	1	2	16	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	75337	75337	2.19	1.48	0	1	2	3	41	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	75337	75337	0.6	0.79	0	0	0	1	5	▇▅▁▂▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	75337	75337	78.77	60.33	0	15	85	116.1	5400	▇▁▁▁▁▁▁▁
Babies	0	75337	75337	0.0048	0.078	0	0	0	0	10	▇▁▁▁▁▁▁▁
Children	0	75337	75337	0.096	0.38	0	0	0	0	3	▇▁▁▁▁▁▁▁

C1 PMS dataset (different period with additional features)

Observations: 52871

Variables: 40

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	52871	52871	2016-08-01	2017-11-28	2017-03-21	485

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	6065	46806	52871	188	9: 23408, NA: 6065, 14: 3001, 7: 2846	FALSE
ArrivalDateMonth	0	52871	52871	12	Oct: 9001, Sep: 7645, Aug: 5524, Nov: 4651	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
AssignedRoomType	0	52871	52871	9	A: 37194, D: 10739, E: 1773, F: 1433	FALSE
AssociatedToEvent	0	52871	52871	2	0: 50611, 1: 2260, NA: 0	FALSE
BookedSPA	0	52871	52871	2	0: 52822, 1: 49, NA: 0	FALSE
Company	50448	2423	52871	178	NA: 50448, 40: 522, 153: 188, 45: 128	FALSE
Country	9	52862	52871	158	PRT: 15947, FRA: 6075, DEU: 5126, GBR: 4649	FALSE
CustomerType	0	52871	52871	4	Tra: 43628, Tra: 8813, Gro: 379, Con: 51	FALSE
DepositType	0	52871	52871	3	No : 46252, Non: 6589, Ref: 30, NA: 0	FALSE
DistributionChannel	0	52871	52871	4	TA/: 45370, Dir: 5206, Cor: 2059, GDS: 236	FALSE
IsCanceled	0	52871	52871	2	0: 31744, 1: 21127, NA: 0	FALSE
IsRepeatedGuest	0	52871	52871	2	0: 51639, 1: 1232, NA: 0	FALSE
MarketSegment	0	52871	52871	7	Onl: 29089, Gro: 8655, Off: 8327, Dir: 4611	FALSE
Meal	0	52871	52871	4	BB: 40634, SC: 9156, HB: 3076, FB: 5	FALSE
RateCode	0	52871	52871	36	OD: 32033, FR: 13557, WA: 982, A4: 931	FALSE
ReservationStatus	0	52871	52871	3	O: 31744, A: 20617, N: 510, NA: 0	FALSE
ReservedRoomType	0	52871	52871	8	A: 39910, D: 9257, E: 1525, F: 1300	FALSE
SRDoubleBed	0	52871	52871	2	0: 37605, 1: 15266, NA: 0	FALSE
SRHighFloor	0	52871	52871	2	0: 50759, 1: 2112, NA: 0	FALSE
SRQuietRoom	0	52871	52871	2	0: 48254, 1: 4617, NA: 0	FALSE
SRTogether	0	52871	52871	2	0: 50628, 1: 2243, NA: 0	FALSE
SRTwinBed	0	52871	52871	2	0: 47249, 1: 5622, NA: 0	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	52871	52871	1.85	0.52	0	2	2	2	4	▁▂▁▇▁▁▁▁
ArrivalDate DayOfMonth	0	52871	52871	15.43	8.68	1	8	15	23	31	▇▇▇▇▆▇▇▆
ArrivalDate WeekNumber	0	52871	52871	29.91	13.53	1	19	33	41	53	▁▂▃▃▂▇▇▁
ArrivalDate Year	0	52871	52871	2016.75	0.43	2016	2016	2017	2017	2017	▃▁▁▁▁▁▁▇
Babies	0	52871	52871	0.0048	0.07	0	0	0	0	2	▇▁▁▁▁▁▁▁
BookingChanges	0	52871	52871	0.23	0.67	0	0	0	0	18	▇▁▁▁▁▁▁▁
DaysInWaiting List	0	52871	52871	2.64	30.74	0	0	0	0	480	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
FolioNumber	0	52871	52871	71313.34	18673.3	5471	58519.5	71995	85606.5	106775	▁▁▂▃▇▇▇▅
LeadTime	0	52871	52871	113.22	121.57	0	23	74	168	823	▇▃▂▁▁▁▁▁
PreviousBookingsNotCanceled	0	52871	52871	0.19	2.34	0	0	0	0	72	▇▁▁▁▁▁▁▁
PreviousCancellations	0	52871	52871	0.015	0.22	0	0	0	0	7	▇▁▁▁▁▁▁▁
RequiredCarParkingSpaces	0	52871	52871	0.022	0.15	0	0	0	0	3	▇▁▁▁▁▁▁▁
StaysInWeekendNights	0	52871	52871	0.82	0.87	0	0	1	2	16	▇▁▁▁▁▁▁▁
StaysInWeekNights	0	52871	52871	2.19	1.46	0	1	2	3	41	▇▁▁▁▁▁▁▁
TotalOfSpecialRequests	0	52871	52871	0.66	0.83	0	0	0	1	5	▇▅▁▂▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	52871	52871	73.91	61.82	0	12	80	117	451.5	▇▆▃▁▁▁▁▁
Children	0	52871	52871	0.094	0.38	0	0	0	0	3	▇▁▁▁▁▁▁▁

C2 PMS dataset

Observations: 32496

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	32496	32496	2016-01-01	2017-12-11	2017-03-03	697

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	32496	32496	126	NUL: 17705, 5: 3517, 238: 1519, 13: 985	FALSE
ArrivalDateMonth	0	32496	32496	12	Sep: 4204, Oct: 3964, May: 3667, Jun: 3231	FALSE
ArrivalDateMonthYear	0	32496	32496	23	201: 2417, 201: 2306, 201: 2019, 201: 1904	FALSE
AssignedRoomType	0	32496	32496	8	A: 18524, B: 9846, J: 1372, C: 1060	FALSE
Company	0	32496	32496	206	NUL: 15879, 33: 6731, 35: 1647, 957: 629	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Country	0	32496	32496	146	PRT: 17984, FRA: 2586, ESP: 1866, BRA: 1073	FALSE
CustomerType	0	32496	32496	3	Tra: 19820, Tra: 12559, Gro: 117, NA: 0	FALSE
DepositType	0	32496	32496	3	No : 25735, Non: 6711, Ref: 50, NA: 0	FALSE
DistributionChannel	0	32496	32496	3	Dir: 12717, Onl: 11818, Con: 7961, NA: 0	FALSE
IsCanceled	0	32496	32496	2	0: 20602, 1: 11894, NA: 0	FALSE
IsRepeatedGuest	0	32496	32496	2	0: 30114, 1: 2382, NA: 0	FALSE
MarketSegment	0	32496	32496	7	Gro: 14154, E-C: 11767, Lei: 2648, Cor: 1841	FALSE
Meal	0	32496	32496	6	BB: 22747, SC: 7157, HB3: 2229, HB4: 150	FALSE
ReservationStatus	0	32496	32496	4	O: 20581, A: 11734, N: 160, R: 21	FALSE
ReservedRoomType	0	32496	32496	8	A: 26713, B: 4034, J: 800, C: 600	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	32496	32496	1.62	0.59	0	1	2	2	28	▇▁▁▁▁▁▁▁
ArrivalDateDay OfMonth	0	32496	32496	15.7	8.77	1	8	16	23	31	▇▇▇▇▆▇▇▆
ArrivalDateWeek Number	0	32496	32496	28.18	12.73	1	18	29	39	53	▁▅▆▇▆▇▆▁
ArrivalDateYear	0	32496	32496	2016.58	0.49	2016	2016	2017	2017	2017	▆▁▁▁▁▁▁▇
BookingChanges	0	32496	32496	0.22	0.87	0	0	0	0	59	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	32496	32496	0.8	10.59	0	0	0	0	308	▇▁▁▁▁▁▁▁
FolioNumber	0	32496	32496	37681	10736	20001	28577	37084	45902	59164	▇▇▇▇▇▆▅▅
LeadTime	0	32496	32496	77.88	104.92	0	9	37	99	910	▇▁▁▁▁▁▁▁
PreviousBookings NotCanceled	0	32496	32496	0.32	2.95	0	0	0	0	82	▇▁▁▁▁▁▁▁
Previous Cancellations	0	32496	32496	0.14	1.18	0	0	0	0	45	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
RequiredCar ParkingSpaces	0	32496	32496	0.00022	0.015	0	0	0	0	1	▇▁▁▁▁▁▁▁
StaysInWeekend Nights	0	32496	32496	0.91	2.46	0	0	1	2	164	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	32496	32496	2.39	5.94	0	1	2	3	409	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	32496	32496	0.008	0.092	0	0	0	0	2	▇▁▁▁▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	32496	32496	65.81	48.26	0	14	74	95	643.72	▇▆▁▁▁▁▁▁
Babies	0	32496	32496	0.0042	0.068	0	0	0	0	2	▇▁▁▁▁▁▁▁
Children	0	32496	32496	0.0082	0.11	0	0	0	0	10	▇▁▁▁▁▁▁▁

C3 PMS dataset

Observations: 14437

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	14437	14437	2016-01-01	2017-11-24	2017-01-25	694

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	14437	14437	38	2: 8367, 4: 2219, 56: 723, NUL: 600	FALSE
ArrivalDateMonth	0	14437	14437	12	Oct: 1566, Sep: 1491, Jun: 1476, May: 1472	FALSE
ArrivalDateMonthYear	0	14437	14437	23	201: 848, 201: 842, 201: 832, 201: 821	FALSE
AssignedRoomType	0	14437	14437	7	C: 5500, A: 5340, B: 1971, E: 788	FALSE
Company	0	14437	14437	71	NUL: 13850, 33: 165, 23: 100, 111: 57	FALSE
Country	0	14437	14437	122	FRA: 2005, DEU: 1435, PRT: 1381, GBR: 1155	FALSE
CustomerType	0	14437	14437	3	Tra: 14090, Tra: 345, Gro: 2, NA: 0	FALSE
DepositType	0	14437	14437	3	No : 12020, Non: 1871, Ref: 546, NA: 0	FALSE
DistributionChannel	0	14437	14437	10	Boo: 8448, Exp: 2239, TA: 1194, OTA: 964	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
IsCanceled	0	14437	14437	2	0: 10582, 1: 3855, NA: 0	FALSE
IsRepeatedGuest	0	14437	14437	2	0: 12843, 1: 1594, NA: 0	FALSE
MarketSegment	0	14437	14437	6	Vac: 13368, Dir: 633, Cor: 284, Com: 62	FALSE
Meal	0	14437	14437	3	BB: 10493, SC: 3943, Und: 1, NA: 0	FALSE
ReservationStatus	0	14437	14437	3	O: 10582, A: 3854, N: 1, NA: 0	FALSE
ReservedRoomType	0	14437	14437	7	A: 8242, C: 4276, B: 1140, E: 365	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	14437	14437	1.78	0.42	0	2	2	2	3	▁▁▂▁▁▇▁▁
ArrivalDateDayOfMonth	0	14437	14437	15.64	8.7	1	8	16	23	31	▇▇▇▇▆▇▇▆
ArrivalDateWeekNumber	0	14437	14437	26.62	13.46	1	16	27	38	53	▅▆▆▇▆▇▆▂
ArrivalDateYear	0	14437	14437	2016.53	0.5	2016	2016	2017	2017	2017	▇▁▁▁▁▁▁▇
BookingChanges	0	14437	14437	0.19	0.58	0	0	0	0	17	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	14437	14437	0.014	0.69	0	0	0	0	55	▇▁▁▁▁▁▁▁
FolioNumber	0	14437	14437	12889.85	4244.84	612	9273	12886	16530	20597	▁▁▇▇▇▇▇▆
LeadTime	0	14437	14437	59.35	74.81	0	14	39	75	798	▇▁▁▁▁▁▁▁
PreviousBookingsNotCanceled	0	14437	14437	0.021	0.18	0	0	0	0	6	▇▁▁▁▁▁▁▁
PreviousCancellations	0	14437	14437	0.49	2.29	0	0	0	0	36	▇▁▁▁▁▁▁▁
RequiredCarParkingSpaces	0	14437	14437	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
StaysInWeekendNights	0	14437	14437	0.83	1.2	0	0	1	2	102	▇▁▁▁▁▁▁▁
StaysInWeekNights	0	14437	14437	2.15	2.52	0	1	2	3	255	▇▁▁▁▁▁▁▁
TotalOfSpecialRequests	0	14437	14437	0.088	0.28	0	0	0	0	1	▇▁▁▁▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	14437	14437	69.28	66.95	0	12	63	126	679	▇▅▁▁▁▁▁▁
Babies	0	14437	14437	0.00014	0.012	0	0	0	0	1	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Children	0	14437	14437	0.0033	0.063	0	0	0	0	2	▇▁▁▁▁▁▁▁

C4 PMS dataset

Observations: 25632

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	25632	25632	2016-01-01	2017-11-30	2017-01-10	697

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	25632	25632	115	969: 7792, NUL: 6010, 967: 1550, 540: 939	FALSE
ArrivalDateMonth	0	25632	25632	12	Oct: 2727, May: 2716, Sep: 2477, Apr: 2407	FALSE
ArrivalDateMonthYear	0	25632	25632	23	201: 1525, 201: 1499, 201: 1338, 201: 1309	FALSE
AssignedRoomType	0	25632	25632	11	A: 13704, B: 2748, I: 2664, C: 2640	FALSE
Company	0	25632	25632	222	NUL: 23179, 993: 205, 103: 196, 123: 147	FALSE
Country	0	25632	25632	129	PRT: 15856, ESP: 1526, FRA: 1256, BRA: 1038	FALSE
CustomerType	0	25632	25632	2	Tra: 22455, Tra: 3177, NA: 0	FALSE
DepositType	0	25632	25632	3	No : 22738, Non: 2460, Ref: 434, NA: 0	FALSE
DistributionChannel	0	25632	25632	4	Und: 23815, TA/: 1782, GDS: 30, Soc: 5	FALSE
IsCanceled	0	25632	25632	2	0: 18671, 1: 6961, NA: 0	FALSE
IsRepeatedGuest	0	25632	25632	2	0: 22212, 1: 3420, NA: 0	FALSE
MarketSegment	0	25632	25632	4	Onl: 21181, TA/: 2626, Cor: 1417, Dir: 408	FALSE
Meal	0	25632	25632	5	BB: 22111, SC: 3125, HB: 333, FB: 62	FALSE
ReservationStatus	0	25632	25632	4	O: 18667, A: 6679, N: 282, R: 4	FALSE
ReservedRoomType	0	25632	25632	11	A: 15968, B: 3038, C: 2582, I: 1476	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	25632	25632	1.88	1.92	0	1	2	2	55	▇▁▁▁▁▁▁▁
ArrivalDateDay OfMonth	0	25632	25632	15.37	8.71	1	8	15	23	31	▇▇▇▇▆▇▇▆

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ArrivalDateWeek Number	0	25632	25632	25.85	13.7	1	15	25	38	53	▆▇▇▇▆▇▇▂
ArrivalDateYear	0	25632	25632	2016.53	0.5	2016	2016	2017	2017	2017	▇▁▁▁▁▁▁▇
BookingChanges	0	25632	25632	0.34	1.04	0	0	0	0	119	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	25632	25632	0.63	8.24	0	0	0	0	371	▇▁▁▁▁▁▁▁
FolioNumber	0	25632	25632	24034	8090	602	17327	24090	30725	39197	▁▁▆▇▇▇▇▆
LeadTime	0	25632	25632	48.68	74.47	0	3	19	59	497	▇▁▁▁▁▁▁▁
PreviousBookings NotCanceled	0	25632	25632	0.38	2.56	0	0	0	0	60	▇▁▁▁▁▁▁▁
Previous Cancellations	0	25632	25632	0.19	0.78	0	0	0	0	13	▇▁▁▁▁▁▁▁
RequiredCar ParkingSpaces	0	25632	25632	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
StaysInWeekend Nights	0	25632	25632	0.83	2.64	0	0	0	1	141	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	25632	25632	2.18	6.41	0	1	1	3	353	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	25632	25632	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	25632	25632	78.65	2025.72	0	7	50	62.5	122070.44	▇▁▁▁▁▁▁▁
Babies	0	25632	25632	0.0047	0.07	0	0	0	0	2	▇▁▁▁▁▁▁▁
Children	0	25632	25632	0.057	0.29	0	0	0	0	3	▇▁▁▁▁▁▁▁

Holidays calendar dataset

Observations: 34

Variables: 2

Variable type: Date

Variable	Missing	Complete	n	Min	Max	Median	n unique
Date	0	34	34	2016-01-01	2017-12-31	2016-12-31	34

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Description	0	34	34	17	All: 2, Ass: 2, Car: 2, Chr: 2	FALSE

Online prices/inventory dataset

Observations: 4676625

Variables: 13

Variable type: Date

Variable	Missing	Complete	n	Min	Max	Median	n unique
LookupDate	0	4676625	4676625	2016-01-02	2018-10-31	2017-04-24	1034
ObservationDate	0	4676625	4676625	2016-01-01	2017-11-26	2016-10-26	687

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
HotelID	0	4676625	4676625	8	34: 1018522, 3: 692432, 2: 629990, 28: 587272	FALSE
Meal	0	4676625	4676625	4	BB: 2674888, SC: 1441758, HB: 523938, FB: 36041	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
CompSetMaxAvailableRooms	0	4676625	4676625	7.62	3.88	0	5	10	10	10	▂▁▁▁▁▁▁▇
HotelsWithOpenSales	0	4676625	4676625	2.07	1.37	0	1	2	3	5	▅▇▁▇▆▁▅▂
MaxAvailableRooms	0	4676625	4676625	7.07	3.1	1	4	8	10	10	▂▂▁▂▁▂▂▇
MaxOccupation	0	4676625	4676625	2.3	1.12	1	2	2	3	6	▅▇▁▃▂▁▁▁
MealNumber	0	4676625	4676625	1.82	0.65	1	1	2	2	4	▅▁▇▁▁▂▁▁
WorseThan	0	4676625	4676625	1.12	1.3	0	0	1	2	5	▇▅▁▃▂▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
MedianCompSetPrice	0	4676625	4676625	109.51	71.77	0	69.3	104.4	147.2	999	▇▃▁▁▁▁▁▁
MinCompSetPrice	0	4676625	4676625	93.51	69.19	0	51	83.6	129	999	▇▂▁▁▁▁▁▁
MinPrice	0	4676625	4676625	150.42	98.78	1	87	126	182	8122	▇▁▁▁▁▁▁▁

R1 PMS dataset

Observations: 35154

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	35154	35154	2016-01-01	2017-12-12	2016-12-03	703

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	35154	35154	167	240: 12263, NUL: 7377, 250: 2695, 241: 1618	FALSE
ArrivalDateMonth	0	35154	35154	12	Oct: 3625, Apr: 3569, May: 3488, Aug: 3457	FALSE
ArrivalDateMonthYear	0	35154	35154	23	201: 1950, 201: 1827, 201: 1801, 201: 1757	FALSE
AssignedRoomType	0	35154	35154	10	A: 14520, D: 9134, E: 5121, C: 2002	FALSE
Company	0	35154	35154	238	NUL: 32331, 223: 587, 405: 139, 154: 137	FALSE
Country	0	35154	35154	124	PRT: 14035, GBR: 6523, ESP: 3236, IRL: 2011	FALSE
CustomerType	0	35154	35154	4	Tra: 26167, Tra: 7126, Con: 1405, Gro: 456	FALSE
DepositType	0	35154	35154	3	No : 33896, Non: 1038, Ref: 220, NA: 0	FALSE
DistributionChannel	0	35154	35154	4	TA/: 25648, Dir: 6875, Cor: 2630, Und: 1	FALSE
IsCanceled	0	35154	35154	2	0: 25597, 1: 9557, NA: 0	FALSE
IsRepeatedGuest	0	35154	35154	2	0: 33339, 1: 1815, NA: 0	FALSE
MarketSegment	0	35154	35154	7	Onl: 16129, Off: 6054, Dir: 5743, Gro: 5132	FALSE
Meal	0	35154	35154	5	BB: 26804, HB: 6883, Und: 964, FB: 421	FALSE
ReservationStatus	0	35154	35154	3	O: 25597, A: 9301, N: 256, NA: 0	FALSE
ReservedRoomType	0	35154	35154	9	A: 19867, D: 6725, E: 4630, G: 1463	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	35154	35154	1.85	0.45	0	2	2	2	4	▁▂▁▇▁▁▁▁
ArrivalDateDay OfMonth	0	35154	35154	15.84	8.81	1	8	16	24	31	▇▇▇▇▆▇▇▇
ArrivalDateWeek Number	0	35154	35154	25.88	13.81	1	14	25	38	53	▆▇▇▇▇▇▇▃
ArrivalDateYear	0	35154	35154	2016.49	0.5	2016	2016	2016	2017	2017	▇▁▁▁▁▁▁▇
BookingChanges	0	35154	35154	0.32	0.78	0	0	0	0	16	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
DaysInWaitingList	0	35154	35154	0.44	8.05	0	0	0	0	480	▇▁▁▁▁▁▁▁
FolioNumber	0	35154	35154	31529	11051	76	22449	31493	40660	53645	▁▂▇▇▇▇▇▂
LeadTime	0	35154	35154	95.54	100.92	-13	10	58	162	725	▇▃▂▁▁▁▁▁
PreviousBookings NotCanceled	0	35154	35154	0.18	1.15	0	0	0	0	31	▇▁▁▁▁▁▁▁
Previous Cancellations	0	35154	35154	0.0089	0.13	0	0	0	0	5	▇▁▁▁▁▁▁▁
RequiredCar ParkingSpaces	0	35154	35154	0.14	0.36	0	0	0	0	9	▇▁▁▁▁▁▁▁
StaysInWeekend Nights	0	35154	35154	1.19	1.14	0	0	1	2	19	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	35154	35154	3.13	2.45	0	1	3	5	50	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	35154	35154	0.63	0.81	0	0	0	1	5	▇▅▁▂▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	35154	35154	67.32	61.75	0	26	54	86	3728	▇▁▁▁▁▁▁▁
Babies	0	35154	35154	0.013	0.11	0	0	0	0	2	▇▁▁▁▁▁▁▁
Children	0	35154	35154	0.13	0.45	0	0	0	0	3	▇▁▁▁▁▁▁▁

R1 PMS dataset (different period with additional features)

Observations: 23827

Variables: 40

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	23827	23827	2016-08-01	2017-12-12	2017-03-13	490

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	5080	18747	23827	143	240: 8324, NA: 5080, 250: 2012, 241: 1255	FALSE
ArrivalDateMonth	0	23827	23827	12	Oct: 3416, Aug: 3012, Sep: 2749, Nov: 2068	FALSE
AssignedRoomType	0	23827	23827	10	A: 9785, D: 5984, E: 3563, C: 1491	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
AssociatedToEvent	0	23827	23827	2	0: 23441, 1: 386, NA: 0	FALSE
BookedSPA	0	23827	23827	2	0: 23821, 1: 6, NA: 0	FALSE
Company	21778	2049	23827	194	NA: 21778, 223: 408, 405: 126, 185: 96	FALSE
Country	175	23652	23827	113	PRT: 8613, GBR: 4776, ESP: 2169, IRL: 1439	FALSE
CustomerType	0	23827	23827	4	Tra: 17382, Tra: 5017, Con: 1031, Gro: 397	FALSE
DepositType	0	23827	23827	3	No : 23311, Non: 405, Ref: 111, NA: 0	FALSE
DistributionChannel	0	23827	23827	4	TA/: 17257, Dir: 4625, Cor: 1944, Und: 1	FALSE
IsCanceled	0	23827	23827	2	0: 17664, 1: 6163, NA: 0	FALSE
IsRepeatedGuest	0	23827	23827	2	0: 22530, 1: 1297, NA: 0	FALSE
MarketSegment	0	23827	23827	7	Onl: 11374, Dir: 3996, Off: 3747, Gro: 3277	FALSE
Meal	0	23827	23827	5	BB: 18093, HB: 4891, Und: 607, FB: 175	FALSE
RateCode	1	23826	23827	48	OD: 13029, FG: 5487, A4: 949, FR: 546	FALSE
ReservationStatus	0	23827	23827	3	O: 17664, A: 5982, N: 181, NA: 0	FALSE
ReservedRoomType	0	23827	23827	9	A: 13070, D: 4638, E: 3292, G: 993	FALSE
SRDoubleBed	0	23827	23827	2	0: 16421, 1: 7406, NA: 0	FALSE
SRHighFloor	0	23827	23827	2	0: 22137, 1: 1690, NA: 0	FALSE
SRQuietRoom	0	23827	23827	2	0: 22804, 1: 1023, NA: 0	FALSE
SRTogether	0	23827	23827	2	0: 22398, 1: 1429, NA: 0	FALSE
SRTwinBed	0	23827	23827	2	0: 21839, 1: 1988, NA: 0	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	23827	23827	1.85	0.45	0	2	2	2	4	▁▂▁▇▁▁▁▁
ArrivalDate DayOfMonth	0	23827	23827	15.7	8.81	1	8	16	23	31	▇▇▇▇▆▇▇▆
ArrivalDate WeekNumber	0	23827	23827	29.15	14.23	1	17	32	41	53	▅▅▅▅▅▇▇▂
ArrivalDate Year	0	23827	23827	2016.71	0.45	2016	2016	2017	2017	2017	▃▁▁▁▁▁▁▇
Babies	0	23827	23827	0.012	0.11	0	0	0	0	2	▇▁▁▁▁▁▁▁
BookingChanges	0	23827	23827	0.35	0.8	0	0	0	0	13	▇▁▁▁▁▁▁▁
DaysInWaiting List	0	23827	23827	0.58	9.52	0	0	0	0	480	▇▁▁▁▁▁▁▁
FolioNumber	0	23827	23827	36990.7	8792.99	4330	31189.5	37410	43648.5	53645	▁▁▁▂▇▇▇▂

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
LeadTime	0	23827	23827	99.48	105.12	-13	11	59	168	725	▇▃▂▁▁▁▁▁
PreviousBookingsNotCanceled	0	23827	23827	0.2	1.26	0	0	0	0	31	▇▁▁▁▁▁▁▁
Previous Cancellations	0	23827	23827	0.0097	0.14	0	0	0	0	5	▇▁▁▁▁▁▁▁
RequiredCar ParkingSpaces	0	23827	23827	0.14	0.37	0	0	0	0	9	▇▁▁▁▁▁▁▁
StaysInWeekend Nights	0	23827	23827	1.22	1.15	0	0	1	2	16	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	23827	23827	3.2	2.48	0	1	3	5	40	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	23827	23827	0.7	0.85	0	0	0	1	5	▇▅▁▂▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	23827	23827	63.57	66.26	0	14	46.8	84.8	3728	▇▁▁▁▁▁▁▁
Children	0	23827	23827	0.14	0.46	0	0	0	0	3	▇▁▁▁▁▁▁▁

R2 PMS dataset

Observations: 8156

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	8156	8156	2016-01-01	2017-12-14	2017-02-02	699

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	8156	8156	73	3: 1826, 19: 987, 288: 677, 21: 637	FALSE
ArrivalDateMonth	0	8156	8156	12	Aug: 1014, Jul: 946, Sep: 946, Oct: 923	FALSE
ArrivalDateMonthYear	0	8156	8156	23	201: 553, 201: 547, 201: 514, 201: 507	FALSE
AssignedRoomType	0	8156	8156	11	C: 2529, B: 2213, A: 1746, G: 1118	FALSE
Company	0	8156	8156	94	NUL: 7532, 31: 31, 297: 27, 70: 26	FALSE
Country	0	8156	8156	49	GBR: 4887, PRT: 825, DEU: 750, ESP: 353	FALSE

Variable	Missing	Complete	n	n unique	Top counts	Ordered
CustomerType	0	8156	8156	3	Tra: 6100, Gro: 1743, Tra: 313, NA: 0	FALSE
DepositType	0	8156	8156	3	No : 6968, Non: 826, Ref: 362, NA: 0	FALSE
DistributionChannel	0	8156	8156	1	Und: 8156, NA: 0	FALSE
IsCanceled	0	8156	8156	2	0: 6671, 1: 1485, NA: 0	FALSE
IsRepeatedGuest	0	8156	8156	2	0: 6360, 1: 1796, NA: 0	FALSE
MarketSegment	0	8156	8156	8	OTA: 3097, TO: 2722, Dir: 885, Own: 638	FALSE
Meal	0	8156	8156	6	BB: 5219, SC: 2210, HB1: 430, HB2: 213	FALSE
ReservationStatus	0	8156	8156	4	O: 6668, A: 1429, N: 56, R: 3	FALSE
ReservedRoomType	0	8156	8156	12	C: 2881, B: 2206, A: 1785, G: 982	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	8156	8156	2.25	0.82	0	2	2	2	7	▁▁▇▁▁▁▁▁
ArrivalDateDay OfMonth	0	8156	8156	15.47	9.09	1	7.75	15	23	31	▇▇▇▇▃▇▇▇
ArrivalDateWeek Number	0	8156	8156	27.83	12.3	1	18	29	37	53	▂▂▅▇▅▇▅▁
ArrivalDateYear	0	8156	8156	2016.54	0.5	2016	2016	2017	2017	2017	▇▁▁▁▁▁▁▇
BookingChanges	0	8156	8156	0.34	0.76	0	0	0	1	19	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	8156	8156	0.068	1.35	0	0	0	0	31	▇▁▁▁▁▁▁▁
FolioNumber	0	8156	8156	20466	2556	13406	18287	20487	22610	25443	▁▂▇▇▇▇▇▅
LeadTime	0	8156	8156	95.43	84.37	0	23	76	146	591	▇▅▂▁▁▁▁▁
PreviousBookings NotCanceled	0	8156	8156	0.78	3.07	0	0	0	0	62	▇▁▁▁▁▁▁▁
Previous Cancellations	0	8156	8156	0.048	0.25	0	0	0	0	4	▇▁▁▁▁▁▁▁
RequiredCar ParkingSpaces	0	8156	8156	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
StaysInWeekend Nights	0	8156	8156	2.02	2.39	0	1	2	2	56	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	8156	8156	5.14	5.78	0	2	5	5	144	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	8156	8156	0.41	0.64	0	0	0	1	5	▇▃▁▁▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	8156	8156	89.84	84.34	0	20	74.4	138	682.65	▇▃▂▁▁▁▁▁
Babies	0	8156	8156	0.054	0.23	0	0	0	0	2	▇▁▁▁▁▁▁▁
Children	0	8156	8156	0.33	0.72	0	0	0	0	4	▇▁▁▁▁▁▁▁

R3 PMS dataset

Observations: 7908

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	7908	7908	2016-01-03	2017-12-15	2017-01-02	698

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	7908	7908	65	94: 935, 168: 799, 3: 741, 43: 672	FALSE
ArrivalDateMonth	0	7908	7908	12	Aug: 1050, Jul: 970, May: 961, Jun: 950	FALSE
ArrivalDateMonthYear	0	7908	7908	23	201: 549, 201: 501, 201: 499, 201: 492	FALSE
AssignedRoomType	0	7908	7908	9	A: 2929, D: 2218, C: 1272, E: 1021	FALSE
Company	0	7908	7908	66	94: 908, 168: 788, 3: 733, 43: 658	FALSE
Country	0	7908	7908	36	D: 1982, NLD: 1515, PRT: 1320, GBR: 850	FALSE
CustomerType	0	7908	7908	3	Con: 4657, Gro: 3056, Tra: 195, NA: 0	FALSE
DepositType	0	7908	7908	3	No : 6971, Ref: 745, Non: 192, NA: 0	FALSE
DistributionChannel	0	7908	7908	6	Spe: 5119, Dir: 969, Sta: 811, OTA: 447	FALSE
IsCanceled	0	7908	7908	2	0: 6823, 1: 1085, NA: 0	FALSE
IsRepeatedGuest	0	7908	7908	2	0: 6165, 1: 1743, NA: 0	FALSE
MarketSegment	0	7908	7908	1	Und: 7908, NA: 0	FALSE
Meal	0	7908	7908	3	SC: 5656, BB: 1268, HB: 984, NA: 0	FALSE
ReservationStatus	0	7908	7908	4	O: 6815, A: 1084, R: 8, N: 1	FALSE
ReservedRoomType	0	7908	7908	9	A: 2929, C: 1926, D: 1719, F: 728	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	7908	7908	2.5	0.91	0	2	2	3	8	▁▇▂▂▁▁▁▁
ArrivalDateDay OfMonth	0	7908	7908	15.62	8.96	1	8	16	24	31	▇▇▇▇▅▇▇▆
ArrivalDateWeek Number	0	7908	7908	27.8	11.7	1	19	28	37	53	▂▃▆▇▇▇▆▁
ArrivalDateYear	0	7908	7908	2016.51	0.5	2016	2016	2017	2017	2017	▇▁▁▁▁▁▁▇
BookingChanges	0	7908	7908	0.35	0.9	0	0	0	0	13	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	7908	7908	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
FolioNumber	0	7908	7908	71088	2659	64027	68910	71073	73188	76790	▁▂▇▇▇▇▆▃
LeadTime	0	7908	7908	145.25	98.31	0	71.75	138	204	557	▇▇▇▅▂▁▁▁
PreviousBookings NotCanceled	0	7908	7908	0.45	1.33	0	0	0	0	20	▇▁▁▁▁▁▁▁
Previous Cancellations	0	7908	7908	0.14	0.36	0	0	0	0	3	▇▁▁▁▁▁▁▁
RequiredCar ParkingSpaces	0	7908	7908	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
StaysInWeekend Nights	0	7908	7908	2.9	2.52	0	2	2	4	52	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	7908	7908	7.3	6.16	0	5	5	10	128	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	7908	7908	0.035	0.24	0	0	0	0	4	▇▁▁▁▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	7908	7908	81.16	56.97	0	42	66.3	112	403.65	▇▇▃▂▁▁▁▁
Babies	0	7908	7908	0.032	0.18	0	0	0	0	2	▇▁▁▁▁▁▁▁
Children	0	7908	7908	0.39	0.76	0	0	0	0	4	▇▁▁▁▁▁▁▁

R4 PMS dataset

Observations: 8871

Variables: 33

Variable type: date

Variable	Missing	Complete	n	Min	Max	Median	n unique
ReservationStatusDate	0	8871	8871	2016-01-01	2017-12-12	2017-01-01	697

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Agent	0	8871	8871	54	837: 2627, 760: 1566, 783: 1421, 784: 1086	FALSE
ArrivalDateMonth	0	8871	8871	12	Apr: 971, Aug: 969, Jul: 969, Oct: 948	FALSE
ArrivalDateMonthYear	0	8871	8871	23	201: 543, 201: 531, 201: 526, 201: 509	FALSE
AssignedRoomType	0	8871	8871	4	A: 8481, B: 317, C: 54, E: 19	FALSE
Company	0	8871	8871	675	NUL: 7206, 112: 49, 115: 24, 851: 24	FALSE
Country	0	8871	8871	66	PRT: 3797, GBR: 2390, ESP: 873, NLD: 419	FALSE
CustomerType	0	8871	8871	4	Con: 4279, Tra: 2681, Gro: 1624, Tra: 287	FALSE
DepositType	0	8871	8871	2	No : 8730, Non: 141, NA: 0	FALSE
DistributionChannel	0	8871	8871	9	Dir: 2793, Gar: 2619, OTA: 1751, TA: 1678	FALSE
IsCanceled	0	8871	8871	2	0: 7145, 1: 1726, NA: 0	FALSE
IsRepeatedGuest	0	8871	8871	2	0: 8382, 1: 489, NA: 0	FALSE
MarketSegment	0	8871	8871	3	Ind: 8758, Gro: 112, Tim: 1, NA: 0	FALSE
Meal	0	8871	8871	6	SC: 4261, AI: 3778, BB: 485, HB: 202	FALSE
ReservationStatus	0	8871	8871	4	O: 7139, A: 1509, N: 217, R: 6	FALSE
ReservedRoomType	0	8871	8871	4	A: 8704, B: 158, C: 8, E: 1	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Adults	0	8871	8871	2.16	0.71	1	2	2	2	6	▁▇▁▁▁▁▁▁
ArrivalDateDayOfMonth	0	8871	8871	15.57	8.91	1	8	15	23	31	▇▆▇▇▅▆▇▆
ArrivalDateWeekNumber	0	8871	8871	27.8	12.75	1	18	28	38	53	▂▃▇▇▇▇▆▂
ArrivalDateYear	0	8871	8871	2016.52	0.5	2016	2016	2017	2017	2017	▇▁▁▁▁▁▁▇
BookingChanges	0	8871	8871	0.41	1.13	0	0	0	0	21	▇▁▁▁▁▁▁▁
DaysInWaitingList	0	8871	8871	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
FolioNumber	0	8871	8871	41758.91	3070.4	32857	39179.5	41817	44132.5	47893	▁▂▆▇▇▇▇▂
LeadTime	0	8871	8871	94	103.78	0	7	54	155	446	▇▂▂▂▁▁▁▁
PreviousBookingsNotCanceled	0	8871	8871	0.09	0.59	0	0	0	0	21	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
Previous Cancellations	0	8871	8871	0.0057	0.083	0	0	0	0	3	▇▁▁▁▁▁▁▁
RequiredCar ParkingSpaces	0	8871	8871	0	0	0	0	0	0	0	▁▁▁▇▁▁▁▁
StaysInWeekend Nights	0	8871	8871	1.71	2.18	0	0	2	2	58	▇▁▁▁▁▁▁▁
StaysInWeek Nights	0	8871	8871	4.43	5.24	0	2	5	5	142	▇▁▁▁▁▁▁▁
TotalOfSpecial Requests	0	8871	8871	0.39	0.7	0	0	0	1	5	▇▂▁▁▁▁▁▁

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
ADR	0	8871	8871	44.81	71.81	0	0	32	64.26	5000	▇▁▁▁▁▁▁▁
Babies	0	8871	8871	0.03	0.18	0	0	0	0	2	▇▁▁▁▁▁▁▁
Children	0	8871	8871	0.29	0.55	0	0	0	0	4	▇▂▁▁▁▁▁▁

Social media reputation

Observations: 6426

Variables: 7

Variable type: Date

Variable	Missing	Complete	n	Min	Max	Median	n unique
Date	0	6426	6426	2016-01-01	2017-12-14	2016-12-22	714

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
HotelCommonID	0	6426	6426	9	1: 714, 2: 714, 3: 714, 6: 714	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
SUMTotalReviewsOnSite	0	6426	6426	1568.88	1223.91	0	559.25	1408	2030	7408	▇▇▆▁▁▁▁▁
WorseThan	0	6426	6426	2.38	1.64	0	1	2	4	5	▇▂▁▇▃▁▇▃

Variable type: numeric

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
AVGCompSet NormalizedRating	0	6426	6426	78.9	4.59	69.93	75.81	78.66	82.3	93.93	▂▂▇▁▂▂▁▁
AVGNormalizedRating	0	6426	6426	79.96	6.83	62.5	75.5	77.5	86.4	93.75	▁▂▂▇▁▂▇▁

MedianCompSetAVG NormalizedRating	0	6426	6426	78.74	4.88	72.2	75.5	76.8	83.8	93.8	▃▇▁▁▂▃▁▁

Special events calendar dataset

Observations: 154

Variables: 4

Variable type: Date

Variable	Missing	Complete	n	Min	Max	Median	n unique
Date	0	154	154	2016-02-18	2017-11-16	2017-03-29	135

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Designation	0	154	154	64	Rot: 8, Alg: 6, F1 : 6, 8th: 4	FALSE
Location	0	154	154	2	Lis: 116, Por: 38, NA: 0	FALSE
Type	0	154	154	2	Min: 136, Maj: 18, NA: 0	FALSE

Weather forecast

Observations: 14020

Variables: 7

Variable type: Date

Variable	Missing	Complete	n	Min	Max	Median	n unique
LookupDate	0	14020	14020	2016-01-01	2017-12-23	2016-12-30	723
ObservationDate	0	14020	14020	2016-01-01	2017-12-14	2016-12-25	705

Variable type: categorical

Variable	Missing	Complete	n	n unique	Top counts	Ordered
Location	0	14020	14020	2	Por: 7050, Lis: 6970, NA: 0	FALSE

Variable type: integer

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
AvgWindInKph	280	13740	14020	19.33	6.39	0	14	19	24	51	▁▂▇▅▃▁▁▁
MaxTemperatureInCelsius	280	13740	14020	22.55	5.69	8	17	22	27	39	▁▂▇▆▇▆▂▁
ProbabilityOfPrecepitation	280	13740	14020	14.54	20.82	0	0	10	20	100	▇▁▁▁▁▁▁▁

Variable	Missing	Complete	n	Mean	SD	p0	p25	p50	p75	p100	Histogram
QuantityOfPrecipitationInMM	280	13740	14020	1.01	3.39	0	0	0	0	74	▇▁▁▁▁▁▁▁

APPENDIX E – ADDITIONAL DATA SOURCES EXTRACTORS: DATABASES DICTIONARIES AND DATABASE SUMMARY STATISTICS

This appendix presents the dictionaries and diagrams of the databases of the extractors built to collect data from other sources than PMS, employed in the work described by Chapter 4. Summary statistics of those databases are also presented. The databases were built using Microsoft SQL Server (version 2014).

Booking.com online reviews

Database with details of online reviews collected from Booking.com, from the hotels studied and from their competitors.

Tables

Name	Description	Columns	Rows
Concept	List of ratings concepts (e.g. cleaning, value for money, etc.)	3	65
GlobalConfig	Global configurations	4	1
Hotel	List of hotels (subjects and competitors)	3	56
LastVerificationDetail	Record of last check for new reviews for each hotel and language	5	168
ObservationDataByConcept	Ratings observed by concept	4	586287
ObservationReviewDetailTag	Tags associated to each review retrieved	3	85660
Username	List of Booking.com users who published the reviews	3	12527
ObservationReviewDetail	Details of each review retrieved	8	32737
Tag	List of tags (e.g. Couple, Travel with family, etc.)	3	677
Observation	Log of all observations checks for reviews	8	54357
Language	List of languages	4	3
ObservationReviewPage	Metadata of reviews pages extracted per hotel observation	5	14824

Table: BookingOnlineReviews.dbo.Concept

Description: List of ratings concepts (e.g. cleaning, value for money, etc.)

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Designation	varchar	50	

Name	Description	Data type	Max length	Nullable
LanguageID	Language ID	int	4	

Data model

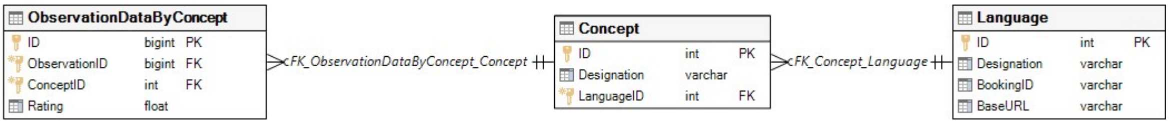


Table: BookingOnlineReviews.dbo.GlobalConfig

Description: Global configurations

Name	Description	Data type	Max length	Nullable
SecondsBetweenHotels	Number of seconds to wait before process another hotel	int	4	
HoursBetweenLastVerification	Interval of time between same hotel check on new reviews (in hours)	int	4	
MinSecondsBetweenPages	Minimum number of seconds to wait for fetching next page of reviews (to use in random selection)	int	4	
MaxSecondsBetweenPages	Maximum number of seconds to wait for fetching next page of reviews (to use in random selection)	int	4	

Data model

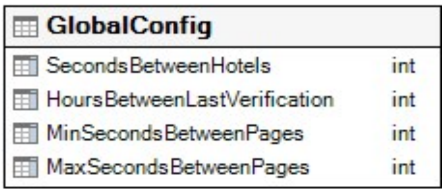

GlobalConfig
SecondsBetweenHotels int
HoursBetweenLastVerification int
MinSecondsBetweenPages int
MaxSecondsBetweenPages int

Table: BookingOnlineReviews.dbo.Hotel

Description: List of hotels (subjects and competitors)

Name	Description	Data type	Max length	Nullable
ID		int	4	
Designation	Hotel name	varchar	50	
URLSuffix	URL of hotel online reviews landing page	varchar	200	

Data model

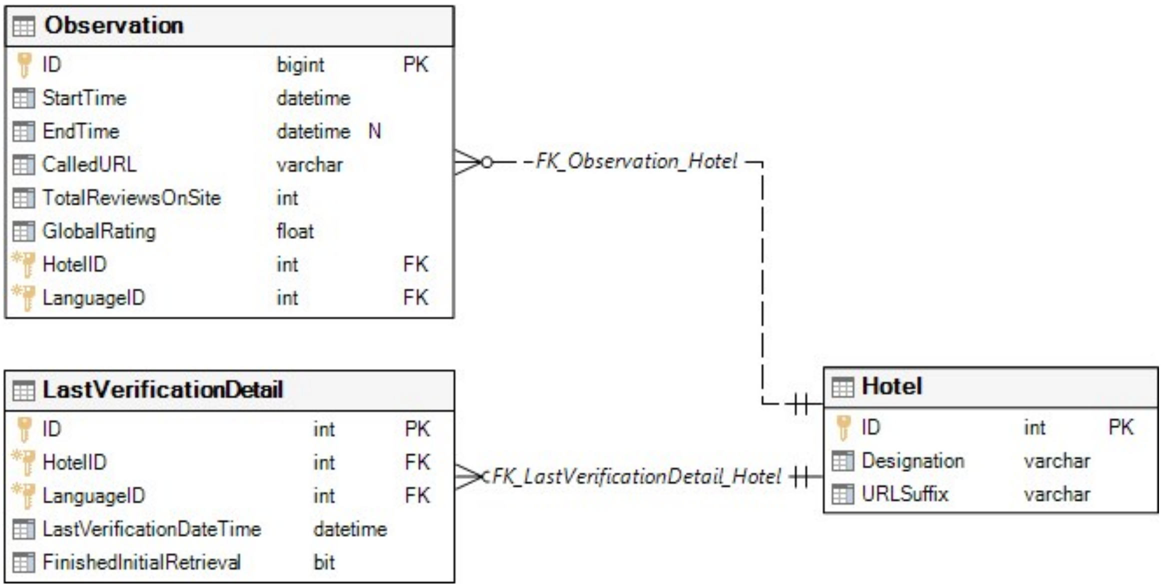


Table: BookingOnlineReviews.dbo.LastVerificationDetail

Description: Record of last check for new reviews for each hotel and language

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
HotelID	Hotel ID	int	4	
LanguageID	Language ID	int	4	
LastVerificationDateTime	Date and time of last check	datetime	8	
FinishedInitialRetrieval	Indication if initial check has finished (when deployed all existing reviews were collected)	bit	1	

Data model

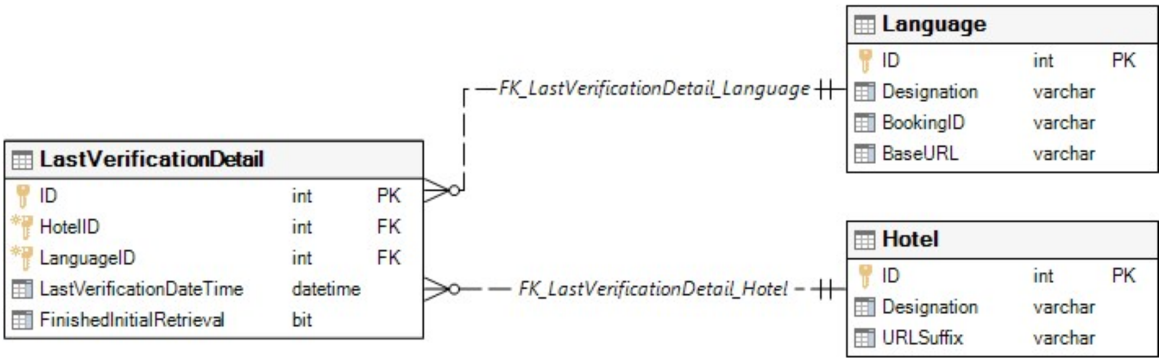


Table: BookingOnlineReviews.dbo.ObservationDataByConcept

Description: Ratings observed by concept

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationID	ID of observation it relates to	bigint	8	
ConceptID	Concept ID	int	4	
Rating	Rating	float	8	

Data model

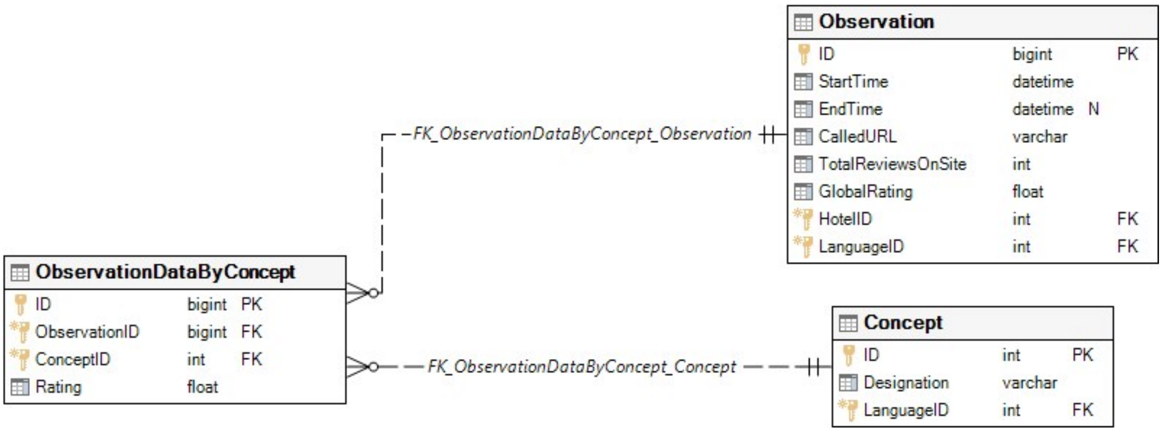


Table: BookingOnlineReviews.dbo.ObservationReviewDetailTag

Description: Tags associated to each review retrieved

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationReviewDetailID	ID of review detail its associated to	bigint	8	
TagID	Tag ID	int	4	

Data model

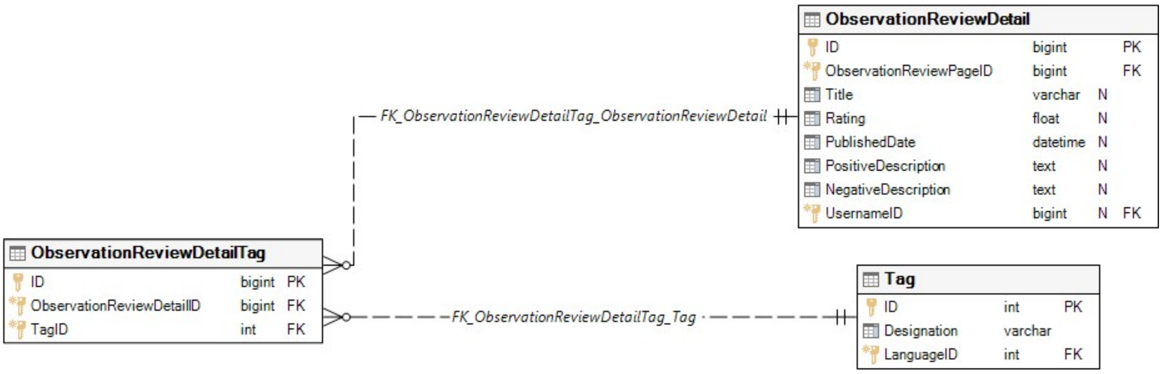

ObservationReviewDetail
ID bigint PK
ObservationReviewPageID bigint FK
Title varchar N
Rating float N
PublishedDate datetime N
PositiveDescription text N
NegativeDescription text N
UsernameID bigint N FK
FK_ObservationReviewDetailTag_ObservationReviewDetail
ObservationReviewDetailTag
ID bigint PK
ObservationReviewDetailID bigint FK
TagID int FK
FK_ObservationReviewDetailTag_Tag
Tag
ID int PK
Designation varchar
LanguageID int FK

Table: BookingOnlineReviews.dbo.Username

Description: List of Booking.com users who published the reviews

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
Name	User name	varchar	100	✓
Location	Location of user	varchar	100	✓

Data model

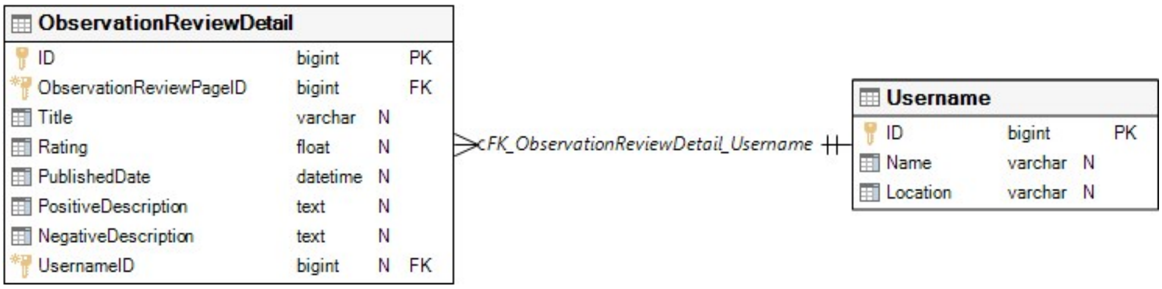


Table: BookingOnlineReviews.dbo.ObservationReviewDetail

Description: Details of each review retrieved

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationReviewPageID	Observation review page ID	bigint	8	
Title	Title of review	varchar	1000	✓
Rating	Overall rating of review	float	8	✓
PublishedDate	Publication date	datetime	8	✓
PositiveDescription	Positive description	text	2147483647	✓
NegativeDescription	Negative description	text	2147483647	✓
UsernameID	ID of Booking.com user	bigint	8	✓

Data model

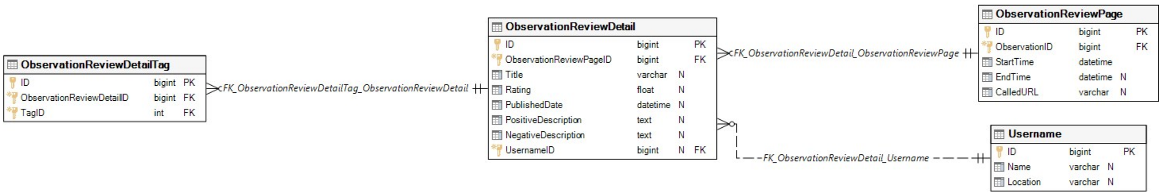


Table: BookingOnlineReviews.dbo.Tag

Description: List of tags (e.g. Couple, Travel with family, etc.)

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Designation	varchar	100	
LanguageID	Language ID	int	4	

Data model

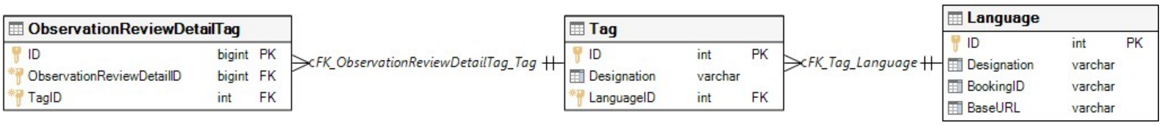


Table: BookingOnlineReviews.dbo.Observation

Description: Log of all observations checks for reviews

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
StartTime	Date and time when operation started	datetime	8	
EndTime	Date and time when operation ended	datetime	8	✓
CalledURL	Full URL of scraped web page	varchar	200	
TotalReviewsOnSite	Number of total reviews on website (independently of the language)	int	4	
GlobalRating	Global rating (independently of the language)	float	8	
HotelID	Hotel ID	int	4	
LanguageID	Language ID	int	4	

Data model

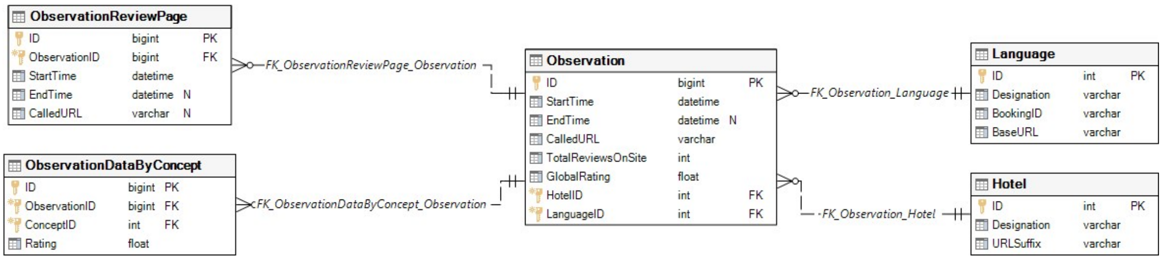


Table: BookingOnlineReviews.dbo.Language

Description: List of languages

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Designation	varchar	20	
BookingID	Booking.com language ID	varchar	6	
BaseURL	Base URL for reviews in the record language	varchar	100	

Data model

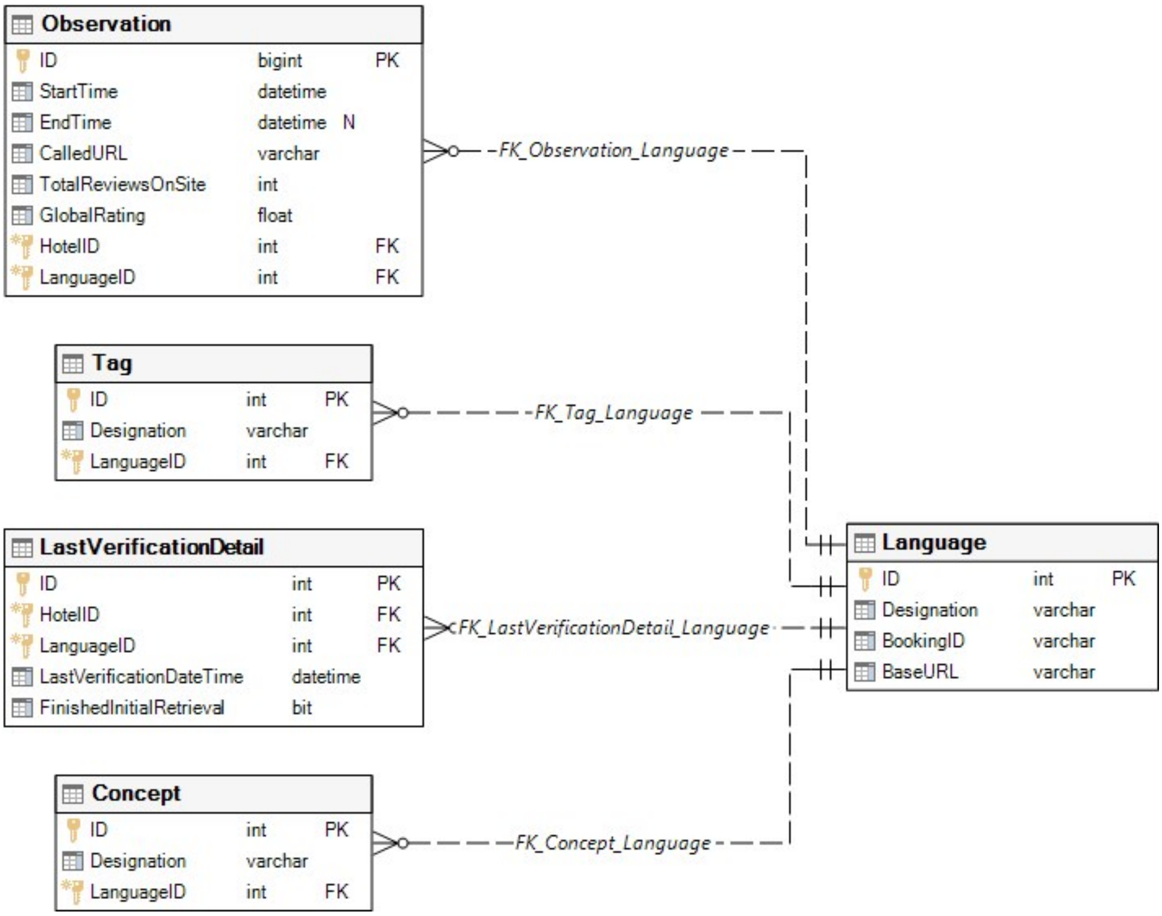
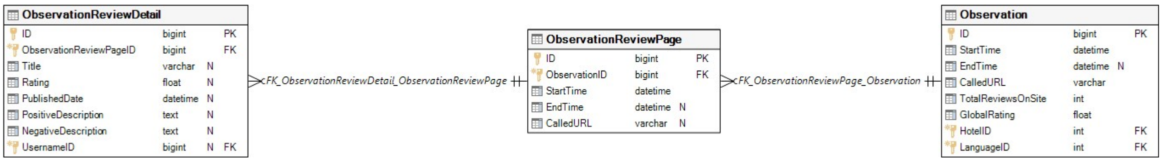


Table: BookingOnlineReviews.dbo.ObservationReviewPage

Description: Metadata of reviews pages extracted per hotel observation

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationID	ID of observation	bigint	8	
StartTime	Start date and time	datetime	8	
EndTime	End date and time	datetime	8	✓
CalledURL	Full scraped web page URL	varchar	2000	✓

Data model



Tripadvisor.com online reviews

Database with details of online reviews collected from Tripadvisor.com, from the hotels studied and from their competitors.

Tables

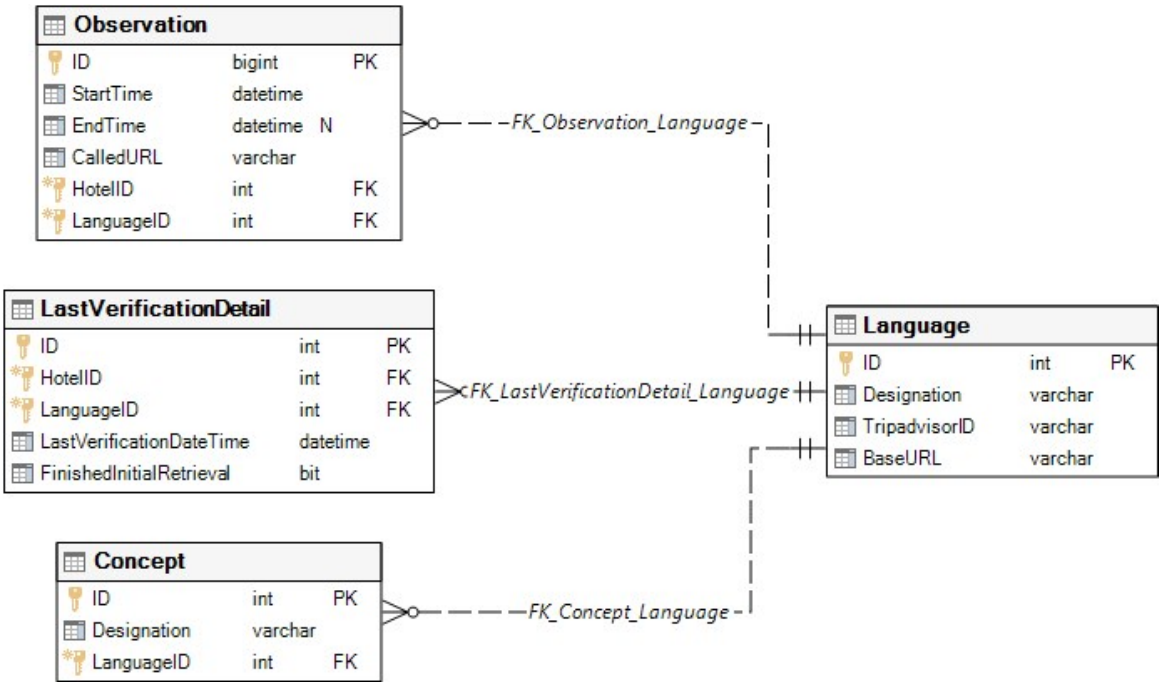
Name	Description	Columns	Rows
Language	List of languages	4	3
Hotel	List of hotels (subjects and competitors)	3	56
GlobalConfig	Global configurations	4	1
Concept	List of concepts (e.g. Value for money, Cleaning, etc.)	3	18
LastVerificationDetail	Maximum number of seconds to wait for fetching next page of reviews (to use in random selection)	5	168
ObservationData	Observations additional data	6	54555
ObservationDataByConcept	Ratings observed by concept	4	65528
ObservationReviewPage	Metadata of reviews pages extracted per hotel observation	5	13491
ObservationReviewDetailConcept	Concept details per review retrieved	4	10571
ObservationReviewDetail	Details of each review retrieved	8	20956
Observation	Log of all observations checks for reviews	6	54555
Username	List of Tripadvisor.com users who published the reviews	3	20920

Table: TripAdvisorOnlineReviews.dbo.Language

Description: List of languages

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Designation	varchar	20	
TripadvisorID	Tripadvisor.com language ID	varchar	3	
BaseURL	Base URL of pages in the record language	varchar	100	

Data model



Observation		
ID	bigint	PK
StartTime	datetime	
EndTime	datetime N	
CalledURL	varchar	
HotelID	int	FK
LanguageID	int	FK

LastVerificationDetail		
ID	int	PK
HotelID	int	FK
LanguageID	int	FK
LastVerificationDateTime	datetime	
FinishedInitialRetrieval	bit	

Concept		
ID	int	PK
Designation	varchar	
LanguageID	int	FK

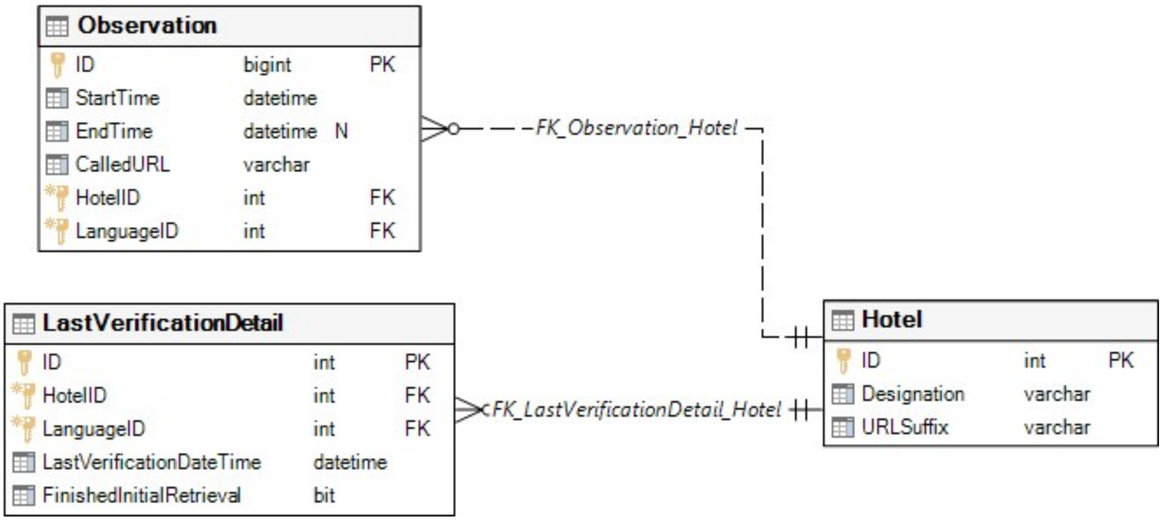
Language		
ID	int	PK
Designation	varchar	
TripadvisorID	varchar	
BaseURL	varchar	

Table: TripAdvisorOnlineReviews.dbo.Hotel

Description: List of hotels (subjects and competitors)

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Hotel name	varchar	50	
URLSuffix	URL of hotel reviews landing page on Tripadvisor.com	varchar	200	

Data model



Observation		
ID	bigint	PK
StartTime	datetime	
EndTime	datetime N	
CalledURL	varchar	
HotelID	int	FK
LanguageID	int	FK

LastVerificationDetail		
ID	int	PK
HotelID	int	FK
LanguageID	int	FK
LastVerificationDateTime	datetime	
FinishedInitialRetrieval	bit	

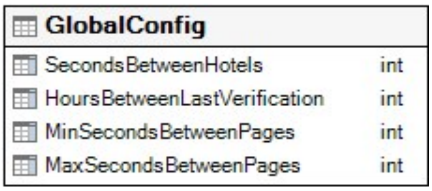
Hotel		
ID	int	PK
Designation	varchar	
URLSuffix	varchar	

Table: TripAdvisorOnlineReviews.dbo.GlobalConfig

Description: Global configurations

Name	Description	Data type	Max length	Nullable
SecondsBetweenHotels	Number of seconds to wait before process another hotel	int	4	
HoursBetweenLastVerification	Interval of time between same hotel check on new reviews (in hours)	int	4	
MinSecondsBetweenPages	Minimum number of seconds to wait for fetching next page of reviews (to use in random selection)	int	4	
MaxSecondsBetweenPages	Maximum number of seconds to wait for fetching next page of reviews (to use in random selection)	int	4	

Data model



GlobalConfig
SecondsBetweenHotels int
HoursBetweenLastVerification int
MinSecondsBetweenPages int
MaxSecondsBetweenPages int

Table: TripAdvisorOnlineReviews.dbo.Concept

Description: List of concepts (e.g. Value for money, Cleaning, etc.)

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Designation	varchar	50	
LanguageID	Language ID	int	4	

Data model

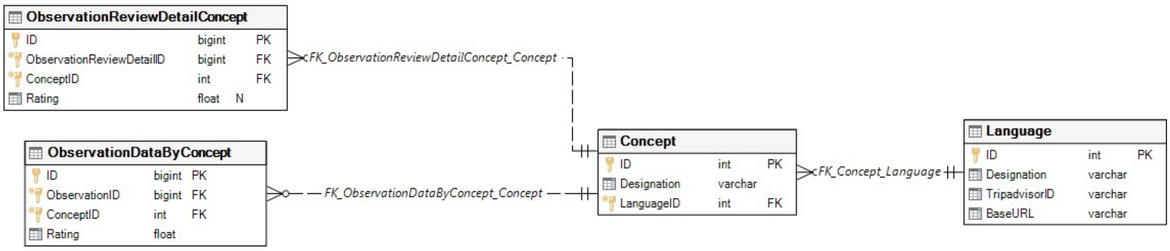


Table: TripAdvisorOnlineReviews.dbo.LastVerificationDetail

Description: Maximum number of seconds to wait for fetching next page of reviews (to use in random selection)

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
HotelID	Hotel ID	int	4	
LanguageID	Language ID	int	4	
LastVerificationDateTime	Date and time of last check	datetime	8	
FinishedInitialRetrieval	Indication if initial check has finished (when deployed all existing reviews were collected)	bit	1	

Data model

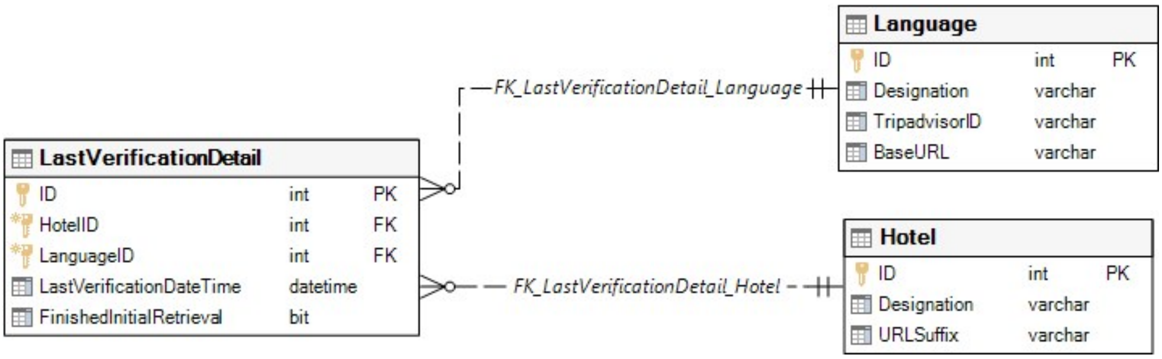


Table: TripAdvisorOnlineReviews.dbo.ObservationData

Description: Observations additional data

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationID	Observation ID it relates to	bigint	8	
TotalReviewsOnSite	Number of total reviews on website (independently of the language)	int	4	
GlobalRating	Global rating (independently of the language)	float	8	
RegionRanking	Ranking in hotel region	int	4	
RegionTotalHotels	Total number of hotels in region	int	4	

Data model

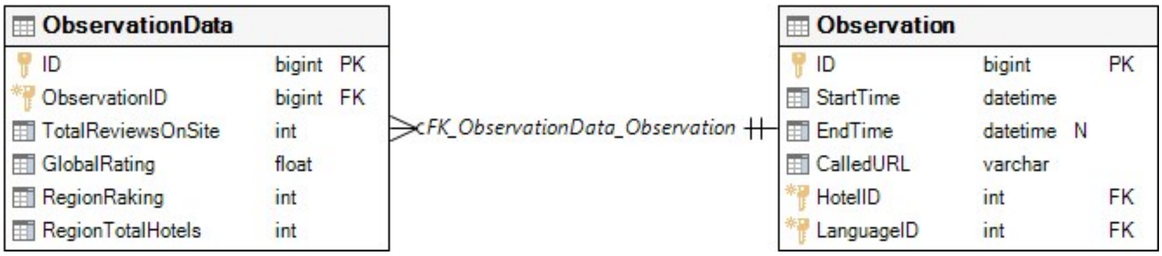


Table: TripAdvisorOnlineReviews.dbo.ObservationDataByConcept

Description: Ratings observed by concept

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationID	ID of observation it relates to	bigint	8	
ConceptID	Concept ID	int	4	
Rating	Rating	float	8	

Data model

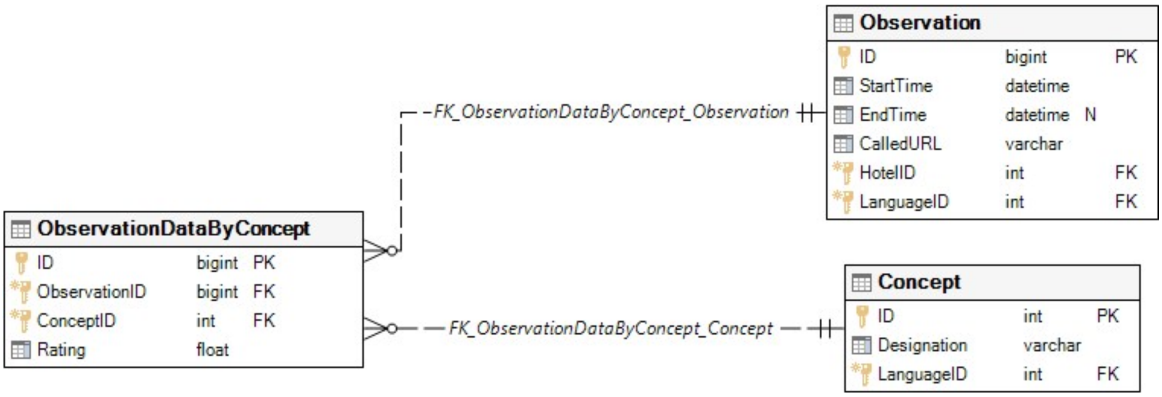


Table: TripAdvisorOnlineReviews.dbo.ObservationReviewPage

Description: Metadata of reviews pages extracted per hotel observation

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationID	ID of observation	bigint	8	
StartTime	Start date and time	datetime	8	
EndTime	End date and time	datetime	8	✓
CalledURL	Full scraped web page URL	varchar	200	

Data model

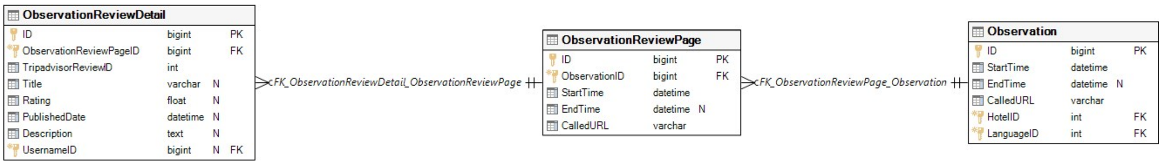


Table: TripAdvisorOnlineReviews.dbo.ObservationReviewDetailConcept

Description: Concept details per review retrieved

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationReviewDetailID	ID of observation review detail	bigint	8	
ConceptID	Concept ID	int	4	
Rating	Rating per concept	float	8	✓

Data model

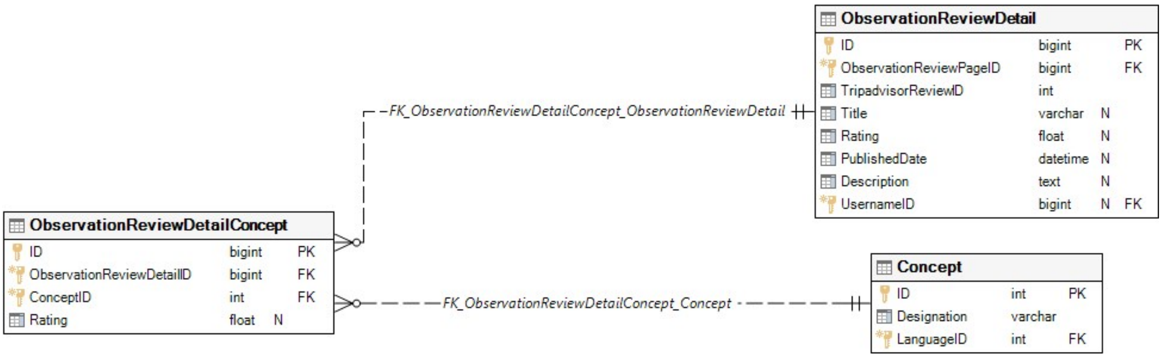


Table: TripAdvisorOnlineReviews.dbo.ObservationReviewDetail

Description: Details of each review retrieved

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationReviewPageID	Observation review page ID	bigint	8	
TripadvisorReviewID	ID of review in Tripadvisor.com	int	4	
Title	Title of review	varchar	1000	✓
Rating	Overall rating of review	float	8	✓
PublishedDate	Publication date	datetime	8	✓
Description	Textual review	text	2147483647	✓
UsernameID	ID of Tripadvisor.com user	bigint	8	✓

Data model

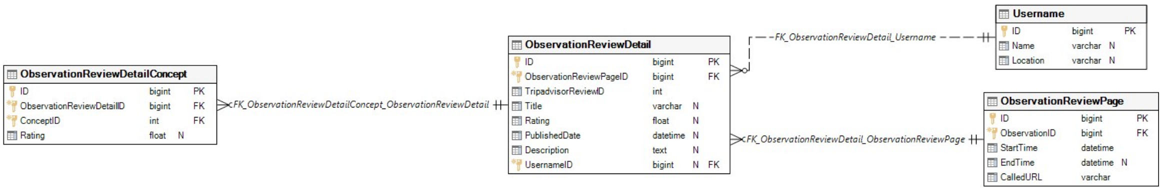


Table: TripAdvisorOnlineReviews.dbo.Observation

Description: Log of all observations checks for reviews

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
StartTime	Date and time when operation started	datetime	8	
EndTime	Date and time when operation ended	datetime	8	✓
CalledURL	Full URL of scraped web page	varchar	200	
HotelID	Hotel ID	int	4	
LanguageID	Language ID	int	4	

Data model

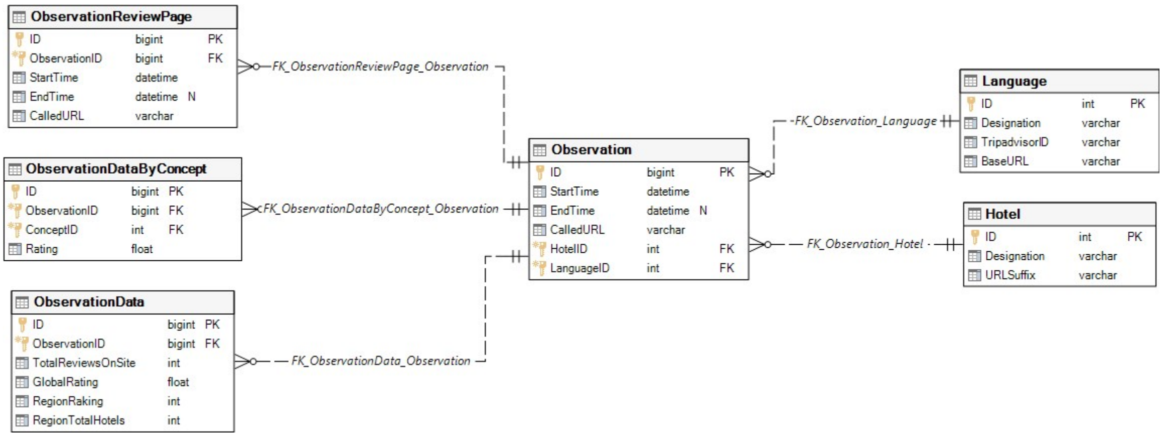
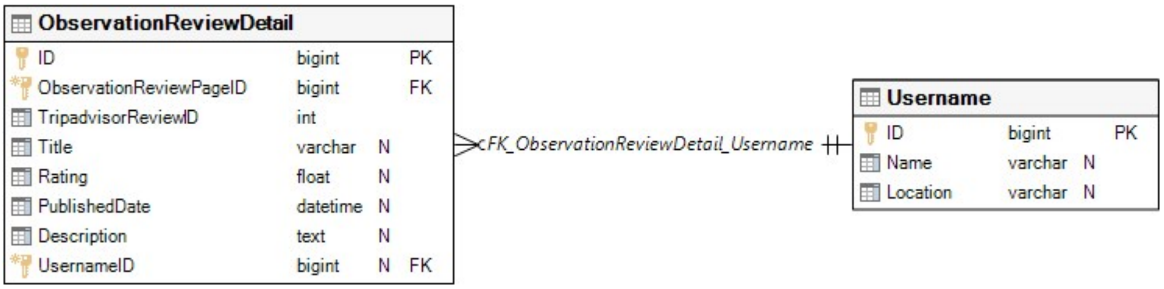


Table: TripAdvisorOnlineReviews.dbo.Username

Description: List of Tripadvisor.com users who published the reviews

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
Name	User name	varchar	100	✓
Location	Location of user	varchar	100	✓

Data model



Booking.com prices and inventory availability

Database with details of prices and rooms on sale for future dates, from studied hotels and from their competitors, collected from Booking.com.

Tables

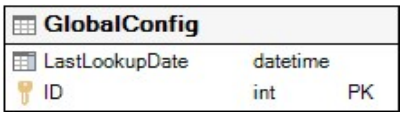
Name	Description	Columns	Rows
GlobalConfig	Global configurations	2	1
Observation	Log of all prices and inventory availability observations made	5	16367522
ObservationDetail	Details of observations	5	89839826
PreProcObservation	Observation data. Observation data is queued in this table for details processing according to computational power availability	8	27308
Hotel	List of hotels to look for information (subjects and competitors)	5	45
RoomType	Room types list	5	4601
BestPricePerHotel	Best price and inventory availability, per day, per hotel, per room capacity and meal type. Processed from observation details.	7	469594

Table: BookingPrices.dbo.GlobalConfig

Description: Global configurations

Name	Description	Data type	Max length	Nullable
LastLookupDate	Date of last day when prices and inventory availability was checked for all hotels	datetime	8	
ID	Internal ID	int	4	

Data model



GlobalConfig		
LastLookupDate	datetime	
ID	int	PK

Table: BookingPrices.dbo.Observation

Description: Log of all prices and inventory availability observations made

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationDateTime	Observation date	datetime	8	
HotelID	Hotel ID	int	4	
LookupDate	Lookup date	datetime	8	
FullObservationDateTime	Observation date and time	datetime	8	✓

Data model

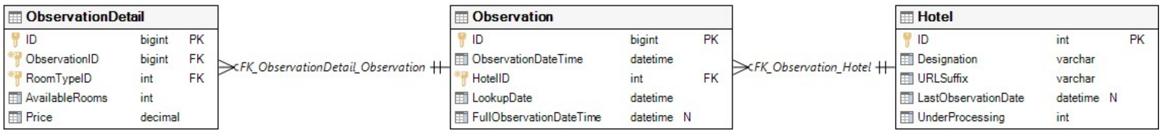


Table: BookingPrices.dbo.ObservationDetail

Description: Details of observations

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ObservationID	Observation ID	bigint	8	
RoomTypeID	Room type ID	int	4	
AvailableRooms	Number of available rooms	int	4	
Price	Price (in EUR)	decimal	9	

Data model

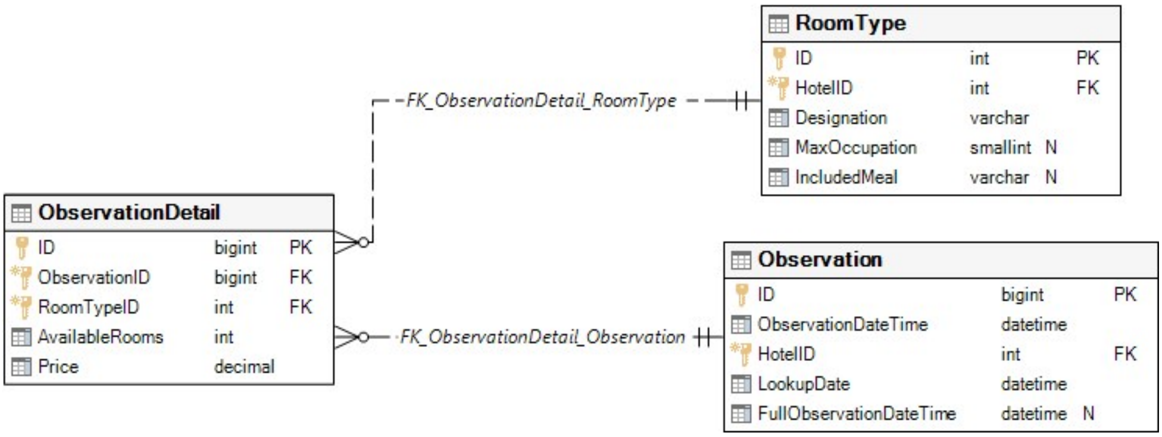


Table: BookingPrices.dbo.PreProcObservation

Description: Observation data. Observation data is queued in this table for details processing according to computational power availability

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
FullObservationDateTime	Data and time of observation	datetime	8	
HotelID	Hotel ID	int	4	
LookupDate	Lookup date	datetime	8	
CalledURL	Full URL of web page scraped	varchar	200	
FullWebPage	Full HTML of web page scraped	text	2147483647	✓
Status	Indication if data processing status (0: to process; 1: under processing, 2: time-out; 9: processed)	int	4	
ProcessStartTime	Date and time when data on the page started to being processed. Used in conjunction to status in distributed computation	datetime	8	✓

Data model

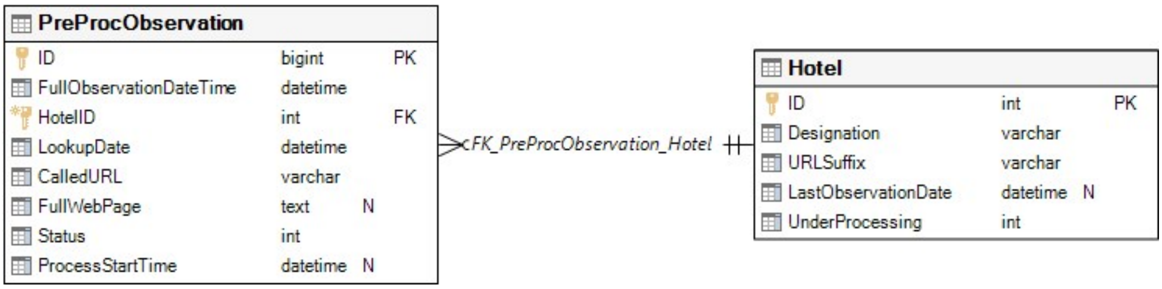


Table: BookingPrices.dbo.Hotel

Description: List of hotels to look for information (subjects and competitors)

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Hotel name	varchar	50	
URLSuffix	Hotel Booking.com prices page URL	varchar	200	
LastObservationDate	Date and time of last time the hotel prices and inventory availability was checked for	datetime	8	✓
UnderProcessing	Indication if hotel is being processed (for distributed computing management)	int	4	

Data model

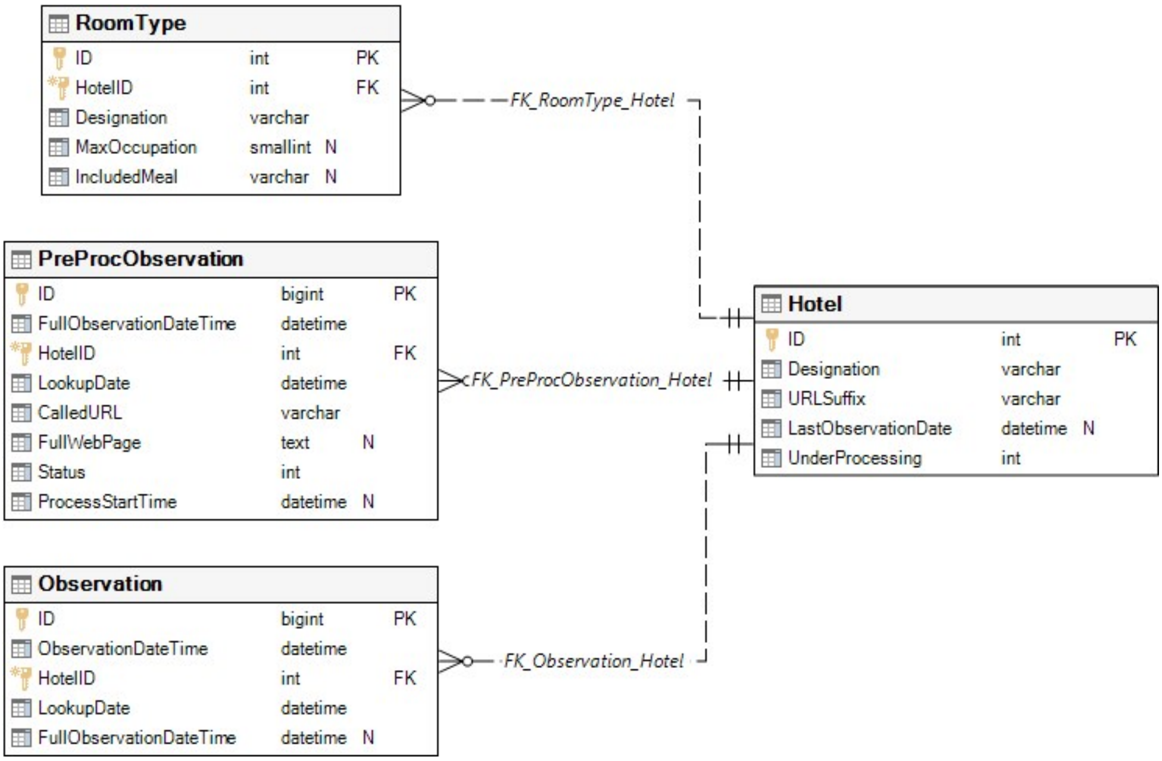


Table: BookingPrices.dbo.RoomType

Description: Room types list

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
HotelID	Hotel ID	int	4	
Designation	Room type designation	varchar	150	
MaxOccupation	Number maximum of adults room type as capacity to	smallint	2	✓
IncludedMeal	Description of meal that is included in the room type	varchar	150	✓

Data model

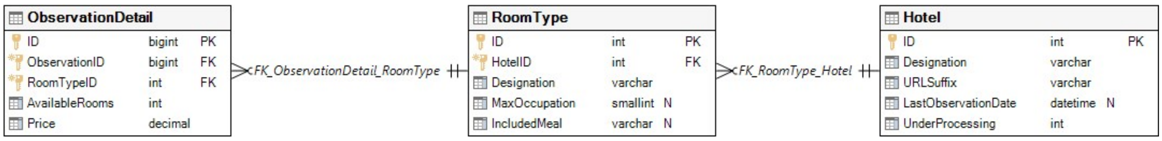
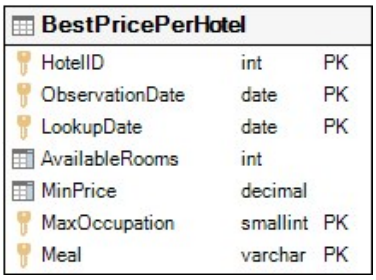


Table: BookingPrices.dbo.BestPricePerHotel

Description: Best price and inventory availability, per day, per hotel, per room capacity and meal type. Processed from observation details.

Name	Description	Data type	Max length	Nullable
HotelID	Hotel ID	int	4	
ObservationDate	Observation date	date	3	
LookupDate	Lookup date	date	3	
AvailableRooms	Number of availabe rooms	int	4	
MinPrice	Minimum price (in EUR)	decimal	9	
MaxOccupation	Maximum adults occupation	smallint	2	
Meal	Meal included in the price	varchar	2	

Data model



BestPricePerHotel

HotelID	int	PK
ObservationDate	date	PK
LookupDate	date	PK
AvailableRooms	int	
MinPrice	decimal	
MaxOccupation	smallint	PK
Meal	varchar	PK

Other data sources

This database includes data from the five remaining data sources, namely: currencies exchange values, events in hotels' regions, holidays per country, stocks exchange indexes and weather forecasts.

Tables

Name	Description	Columns	Rows
Weather	Weather API calls details	18	3796
CurrencyExchange	Currency exchange API calls details	5	1813
WeatherForecast	Details of weather forecast for each lookup date, of each observed date	16	14250
HolidayDetail	Holidays processed from the web page scraping	5	14789
StockIndex	Log of stock exchange index web page scraping	5	1259
Location	Locations of hotels	2	2
EventsScraping	Log of events web page scraping	5	695
EventsScrapingDetail	Events processed from the web page scraping	5	3956
EventsManual	List of special events that happened in the region of the hotels	7	0
CurrencyExchangeDetail	Currency exchange processed results (for querying easier)	4	302404
StockIndexName	Identification of stocks indexes	3	68

Name	Description	Columns	Rows
StockIndexDetail	Stocks exchange indexes processed from the web page scraping	4	73078
Holiday	Log of holidays web page scraping	6	1368
Country	Countries details	6	253
GlobalConfig	Global configurations	8	1

Table: SecondaryDatasources.dbo.Weather

Description: Weather API calls details

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
LocationID	ID of location	int	4	
FullJSONReturn	Full JSON data returned	text	2147483647	✓
CallURL	Full URL of API call	varchar	500	
IsHistory	Indication if is a historic lookup (previous date) or a current date	bit	1	
ObservationDateTime	Date and time of observation	datetime	8	
LookupDate	Date and time of lookup date	datetime	8	
TemperatureInCelsius	Temperature in Celsius degrees	float	8	
RelativeHumidityInPercentage	Percentage of relative humidity	varchar	5	
WindDirection	Direction of the wind	varchar	5	
WindDegrees	Wind direction (in degrees)	smallint	2	
WindInKph	Wind strength in km per hour	float	8	
WindGustInKph	Wind gusts in km per hour	float	8	
UV	Ultra-violet level	float	8	
Precipitation1HourInMeters	Precipitation in last hour (in meters)	float	8	
PrecipitationTodayInMeters	Precipitation of the day (in meters)	float	8	
IconURL	URL of weather icon	varchar	150	
ForecastedPeriods	How many days was forecast retrieved for (default was 10)	tinyint	1	

Data model

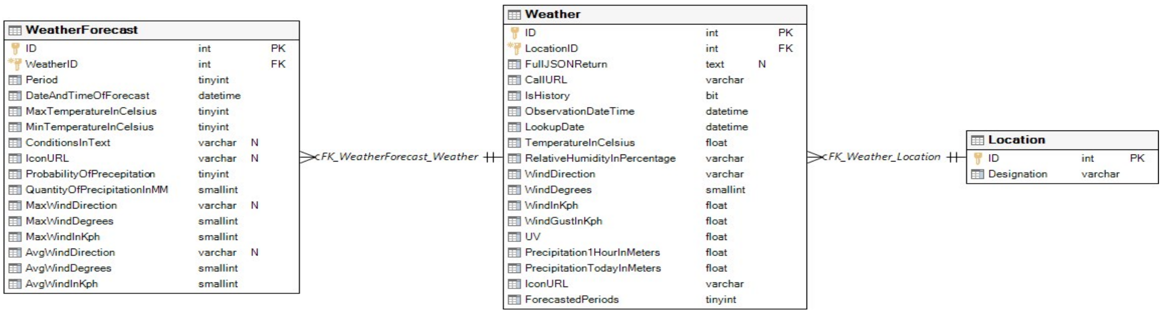

WeatherForecast
ID int PK
WeatherID int FK
Period tinyint
DateAndTimeOfForecast datetime
MaxTemperatureInCelsius tinyint
MinTemperatureInCelsius tinyint
ConditionsInText varchar N
IconURL varchar N
ProbabilityOfPrecepitation tinyint
QuantityOfPrecipitationInMM smallint
MaxWindDirection varchar N
MaxWindDegrees smallint
MaxWindInKph smallint
AvgWindDirection varchar N
AvgWindDegrees smallint
AvgWindInKph smallint
FK_WeatherForecast_Weather
Weather
ID int PK
LocationID int FK
FullJSONReturn text N
CallURL varchar
IsHistory bit
ObservationDateTime datetime
LookupDate datetime
TemperatureInCelsius float
RelativeHumidityInPercentage varchar
WindDirection varchar
WindDegrees smallint
WindInKph float
WindGustInKph float
UV float
Precipitation1HourInMeters float
PrecipitationTodayInMeters float
IconURL varchar
ForecastedPeriods tinyint
FK_Weather_Location
Location
ID int PK
Designation varchar

Table: SecondaryDatasources.dbo.CurrencyExchange

Description: Currency exchange API calls details

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
ObservationDateTime	Observation date and time	datetime	8	
LookupDatetime	Lookup date and time	datetime	8	
FullJSONReturn	Full JSON data returned by API	text	2147483647	✓
CallURL	Full URL API call	varchar	500	

Data model

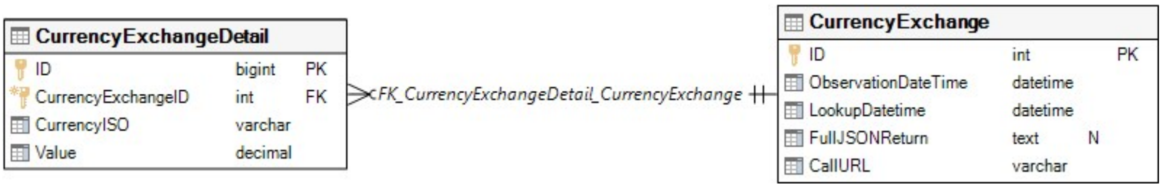


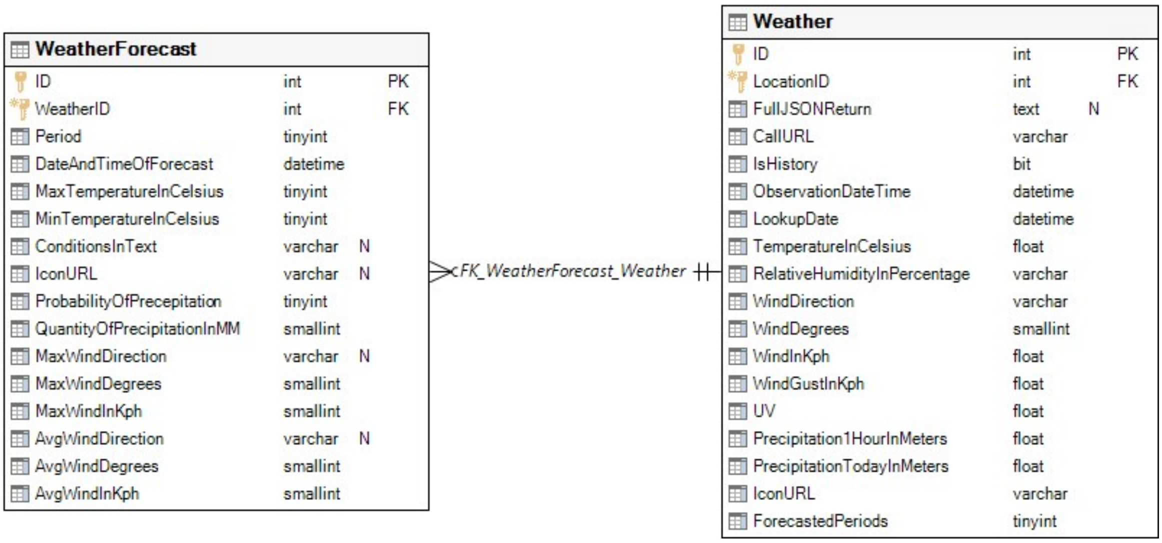
Table: SecondaryDatasources.dbo.WeatherForecast

Description: Details of weather forecast for each lookup date, of each observed date

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
WeatherID	Weather API call ID	int	4	
Period	Period to which the forecast is about (1-10, being 1 the day following the observation date and 10, the 10th day after the observation date)	tinyint	1	
DateAndTimeOfForecast	Date and time of forecast	datetime	8	
MaxTemperatureInCelsius	Maximum temperature forecasted (in Celsius degrees)	tinyint	1	
MinTemperatureInCelsius	Minimum temperature forecasted (in Celsius degrees)	tinyint	1	

Name	Description	Data type	Max length	Nullable
ConditionsInText	Qualitative description of forecast	varchar	50	✓
IconURL	URL of forecast icon	varchar	150	✓
ProbabilityOfPrecipitation	Probability of precipitation in percentage	tinyint	1	
QuantityOfPrecipitationInMM	Prediction of precipitation in mm	smallint	2	
MaxWindDirection	Maximum wind direction	varchar	5	✓
MaxWindDegrees	Maximum wind degrees	smallint	2	
MaxWindInKph	Maximum wind in km per hour	smallint	2	
AvgWindDirection	Average wind direction	varchar	5	✓
AvgWindDegrees	Average wind in degrees	smallint	2	
AvgWindInKph	Average wind in km per hour	smallint	2	

Data model



WeatherForecast			
ID	int		PK
WeatherID	int		FK
Period	tinyint		
DateAndTimeOfForecast	datetime		
MaxTemperatureInCelsius	tinyint		
MinTemperatureInCelsius	tinyint		
ConditionsInText	varchar	N	
IconURL	varchar	N	
ProbabilityOfPrecepitation	tinyint		
QuantityOfPrecipitationInMM	smallint		
MaxWindDirection	varchar	N	
MaxWindDegrees	smallint		
MaxWindInKph	smallint		
AvgWindDirection	varchar	N	
AvgWindDegrees	smallint		
AvgWindInKph	smallint		

Weather			
ID	int		PK
LocationID	int		FK
FullJSONReturn	text	N	
CallURL	varchar		
IsHistory	bit		
ObservationDateTime	datetime		
LookupDate	datetime		
TemperatureInCelsius	float		
RelativeHumidityInPercentage	varchar		
WindDirection	varchar		
WindDegrees	smallint		
WindInKph	float		
WindGustInKph	float		
UV	float		
Precipitation1HourInMeters	float		
PrecipitationTodayInMeters	float		
IconURL	varchar		
ForecastedPeriods	tinyint		

Table: SecondaryDatasources.dbo.HolidayDetail

Description: Holidays processed from the web page scraping

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
HolidayID	ID of holiday log	int	4	
LocationID	Location ID	int	4	✓
Date	Date of the holiday	date	3	
Designation	Holiday designation	varchar	120	

Data model

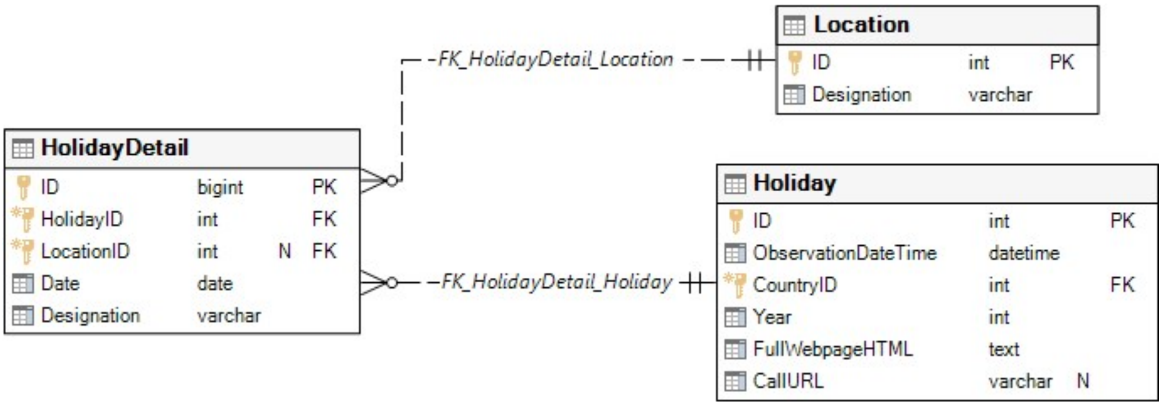


Table: SecondaryDatasources.dbo.StockIndex

Description: Log of stock exchange index web page scraping

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
ObservationDateTime	Date and time of observation	datetime	8	
LookupDatetime	Lookup date and time	datetime	8	
FullWebpageHTML	Full HTML of web page scraped	text	2147483647	✓
CallURL	Full URL of the page	varchar	200	✓

Data model

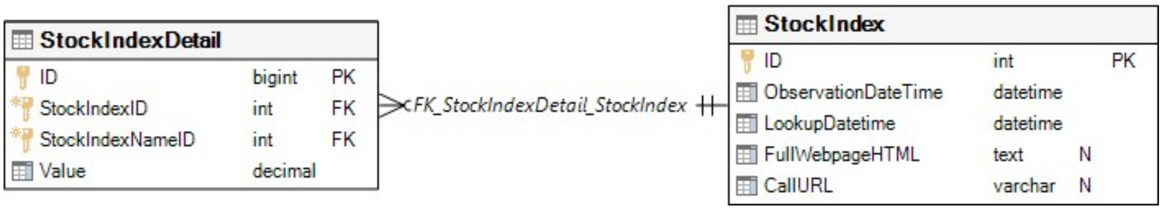
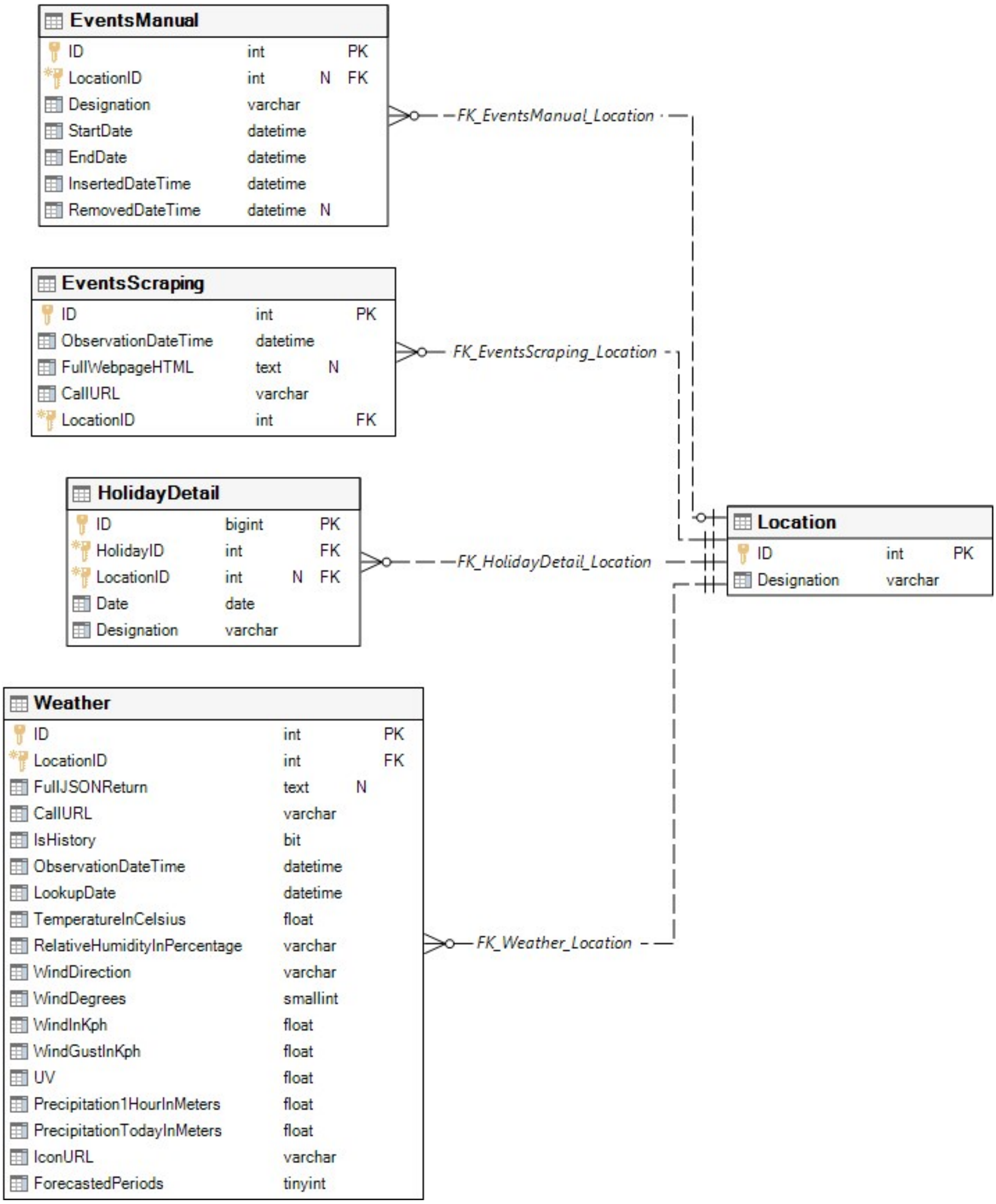


Table: SecondaryDatasources.dbo.Location

Description: Locations of hotels

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Designation	varchar	20	

Data model



EventsManual			
ID	int		PK
LocationID	int	N	FK
Designation	varchar		
StartDate	datetime		
EndDate	datetime		
InsertedDateTime	datetime		
RemovedDateTime	datetime	N	

EventsScraping			
ID	int		PK
ObservationDateTime	datetime		
FullWebpageHTML	text	N	
CallURL	varchar		
LocationID	int		FK

HolidayDetail			
ID	bigint		PK
HolidayID	int		FK
LocationID	int	N	FK
Date	date		
Designation	varchar		

Weather			
ID	int		PK
LocationID	int		FK
FullJSONReturn	text	N	
CallURL	varchar		
IsHistory	bit		
ObservationDateTime	datetime		
LookupDate	datetime		
TemperatureInCelsius	float		
RelativeHumidityInPercentage	varchar		
WindDirection	varchar		
WindDegrees	smallint		
WindInKph	float		
WindGustInKph	float		
UV	float		
Precipitation1HourInMeters	float		
PrecipitationTodayInMeters	float		
IconURL	varchar		
ForecastedPeriods	tinyint		

Location		
ID	int	PK
Designation	varchar	

Table: SecondaryDatasources.dbo.EventsScraping

Description: Log of events web page scraping

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
ObservationDateTime	Date and time of observation	datetime	8	

Name	Description	Data type	Max length	Nullable
FullWebpageHTML	Full HTML of web page	text	2147483647	✓
CallURL	Full URL used of web page scraped	varchar	500	
LocationID	ID of the location	int	4	

Data model

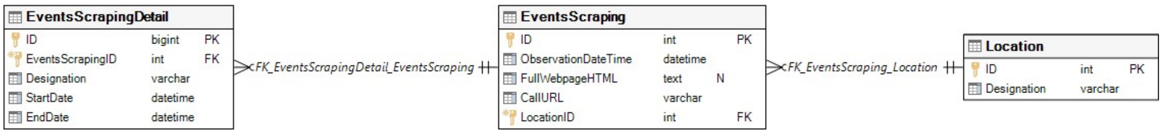


Table: SecondaryDatasources.dbo.EventsScrapingDetail

Description: Events processed from the web page scraping

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
EventsScrapingID	ID of the web scraping log which is associated to	int	4	
Designation	Event designation	varchar	200	
StartDate	Date event starts	datetime	8	
EndDate	Date event ends	datetime	8	

Data model

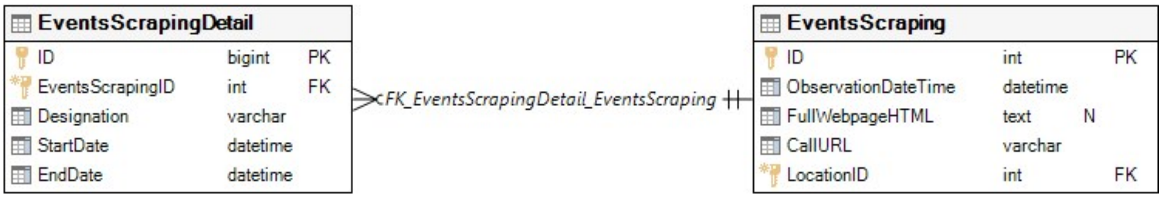


Table: SecondaryDatasources.dbo.EventsManual

Description: List of special events that happened in the region of the hotels

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
LocationID	ID of event location	int	4	✓
Designation	Name of event	varchar	200	
StartDate	Date event starts	datetime	8	
EndDate	Date event ends	datetime	8	
InsertedDateTime	Date and time when event was created in the database (important for knowing if record can be used for processing bookings)	datetime	8	

Name	Description	Data type	Max length	Nullable
RemovedDateTime	Date and time if event was removed (eventually events can be canceled)	datetime	8	✓

Data model

Table: SecondaryDatasources.dbo.CurrencyExchangeDetail

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
CurrencyExchangeID	ID of currency exchange observation	int	4	
CurrencyISO	Currency ISO code	varchar	3	
Value	Currency exchange value to EUR	decimal	9	

Data model

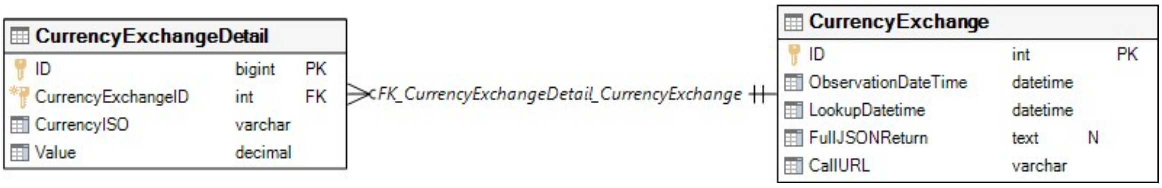


Table: SecondaryDatasources.dbo.StockIndexName

Description: Identification of stocks indexes

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
IndexName	Name of stock index	varchar	50	
CountryID	ID of country associated to the index	int	4	

Data model

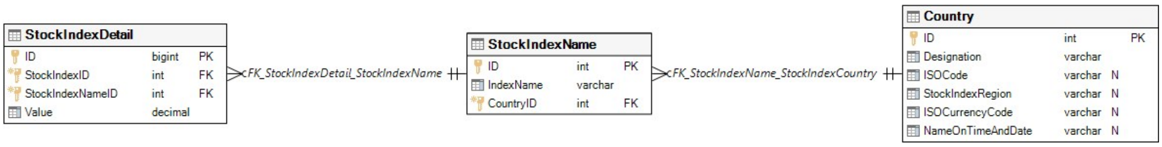


Table: SecondaryDatasources.dbo.StockIndexDetail

Description: Stocks exchange indexes processed from the web page scraping

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
StockIndexID	Stock index observation log ID	int	4	

Name	Description	Data type	Max length	Nullable
StockIndexNameID	Stock index ID	int	4	
Value	Value of the index	decimal	9	

Data model

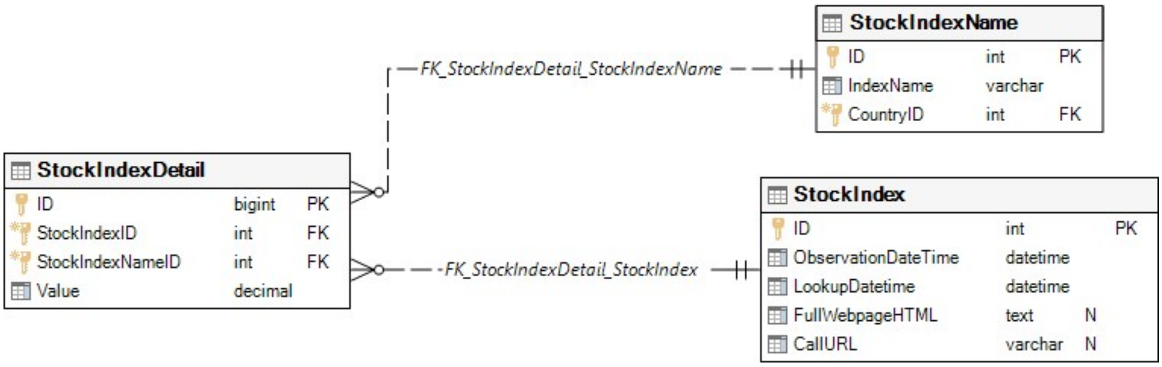


Table: SecondaryDatasources.dbo.Holiday

Description: Log of holidays web page scraping

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
ObservationDateTime	Date and time of observation	datetime	8	
CountryID	Country ID	int	4	
Year	Year the observation is about	int	4	
FullWebpageHTML	Full web page HTML	text	2147483647	
CallURL	Full URL of the web page	varchar	500	✓

Data model

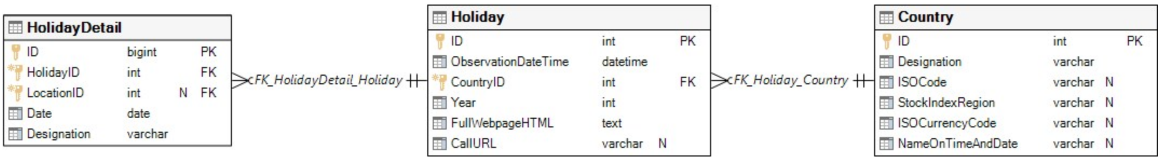


Table: SecondaryDatasources.dbo.Country

Description: Countries details

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
Designation	Designation	varchar	50	
ISOCode	Country ISO code	varchar	3	✓

Name	Description	Data type	Max length	Nullable
StockIndexRegion	Name of main stock exchange index used on the region	varchar	15	✓
ISOCurrencyCode	Local official currency ISO code	varchar	3	✓
NameOnTimeAndDate	Name of the country on TimeAndDate.com	varchar	50	✓

Data model

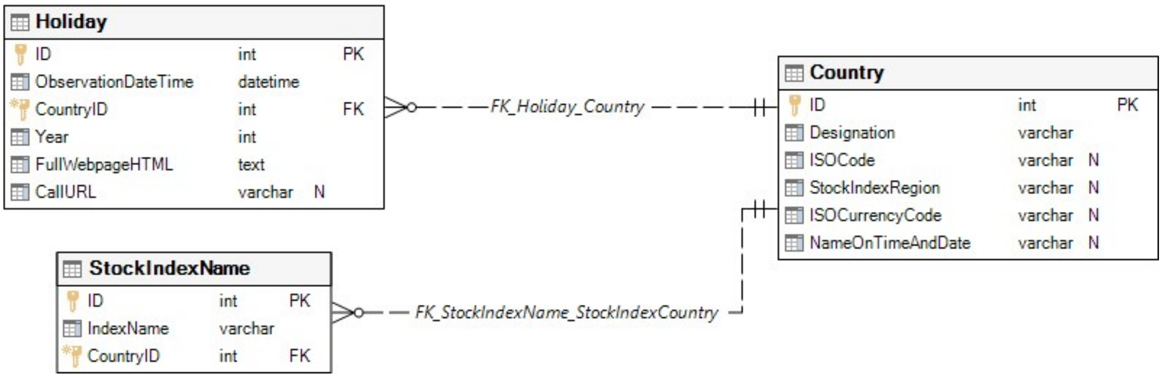
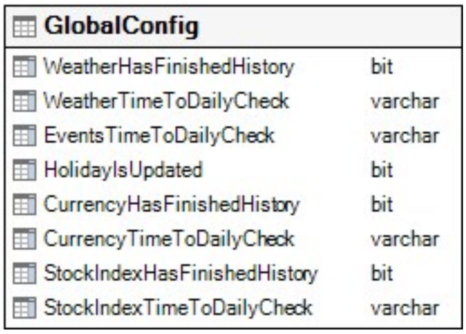


Table: SecondaryDatasources.dbo.GlobalConfig

Description: Global configurations

Name	Description	Data type	Max length	Nullable
WeatherHasFinishedHistory	Indication if historic weather information (dates before extractor was deployed) has finished	bit	1	
WeatherTimeToDailyCheck	Time (hh:mm) to daily check weather forecast	varchar	5	
EventsTimeToDailyCheck	Time (hh:mm) to daily check events on locations	varchar	5	
HolidayIsUpdated	Indication if holidays calendar is updated	bit	1	
CurrencyHasFinishedHistory	Indication if historic currency information (dates before extractor was deployed) has finished	bit	1	
CurrencyTimeToDailyCheck	Time (hh:mm) to daily check currency exchange values	varchar	5	
StockIndexHasFinishedHistory	Indication if historic stock exchange information (dates before extractor was deployed) has finished	bit	1	
StockIndexTimeToDailyCheck	Time (hh:mm) to daily check stock exchange information	varchar	5	

Data model



GlobalConfig	
WeatherHasFinishedHistory	bit
WeatherTimeToDailyCheck	varchar
EventsTimeToDailyCheck	varchar
HolidayIsUpdated	bit
CurrencyHasFinishedHistory	bit
CurrencyTimeToDailyCheck	varchar
StockIndexHasFinishedHistory	bit
StockIndexTimeToDailyCheck	varchar

APPENDIX F – PROTOTYPE DATABASE DICTIONARY AND DATABASE SUMMARY STATISTICS

Prototype database dictionary, diagram and summary statistics as described in Chapter 5 are here presented. Database was built in Microsoft SQL Server (version 2014).

Tables

Name	Description	Columns	Rows
AspNetRoles	User roles	2	2
AspNetUserClaims	User claims logs. Claims represent what users are, not what users can do	4	0
AspNetUserLogins	User logins	3	0
PredictionSummary	Booking's prediction statistics per processing date	10	1570048
AspNetUserRoles	List of roles each user belongs to	2	10
AspNetUsers	List of system users	12	5
Booking	Bookings details	14	118120
ExecutionLog	Log of models' daily processing	8	98
Hotel	List of hotels	6	2
Model	Built models parameters	11	99
ModelPrediction	Models' predictions per execution and booking	4	785024
Performance	Models' execution statistics	15	388
RoomType	Hotels' room types	4	15
SupplyAndDemand	Totals of rooms available for sale in inventory and rooms sold per day	6	1610385
WebsiteLog	Log of operations conducted on the website	6	526
WebsiteUserAction	Log of customers contacts to prevent cancellation	9	17

Table: BCPrototype.dbo.AspNetRoles

Description: User roles

Name	Description	Data type	Max length	Nullable
Id	Internal ID	nvarchar	128	
Name	Role name	nvarchar	256	

Data model

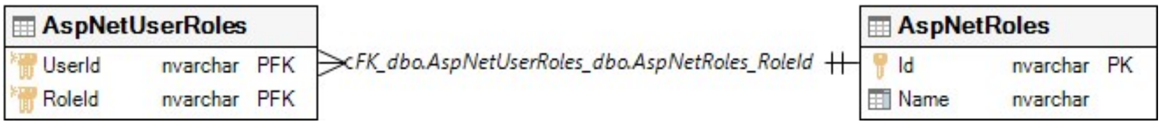


Table: BCPrototype.dbo.AspNetUserClaims

Description: User claims logs. Claims represent what users are, not what users can do

Name	Description	Data type	Max length	Nullable
Id	Internal ID	int	4	
UserId	User ID	nvarchar	128	
ClaimType	Claim type	nvarchar	1073741823	✓
ClaimValue	Claim value	nvarchar	1073741823	✓

Data model

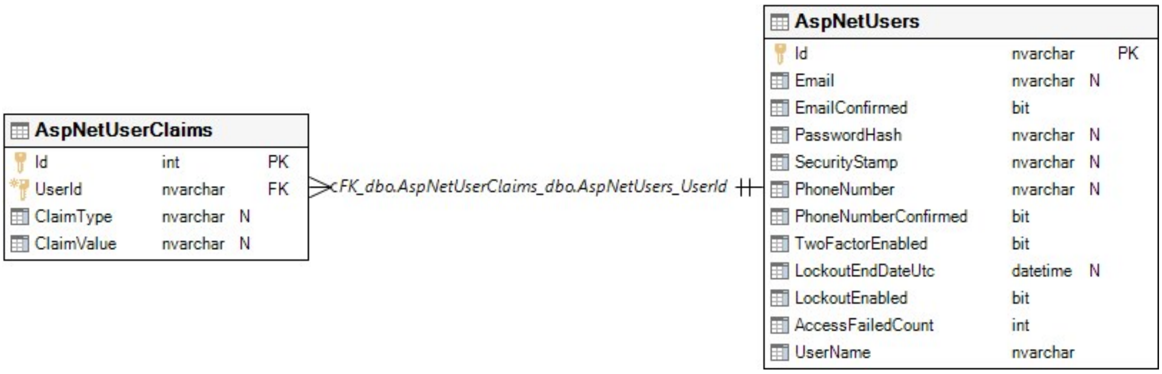


Table: BCPrototype.dbo.AspNetUserLogins

Description: User logins

Name	Description	Data type	Max length	Nullable
LoginProvider	Login provider	nvarchar	128	
ProviderKey	Provider key	nvarchar	128	
UserId	User ID	nvarchar	128	

Data model

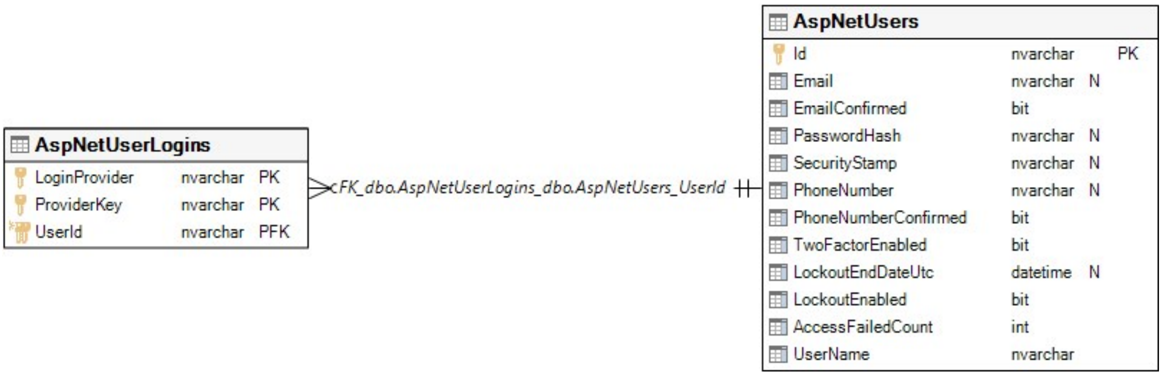
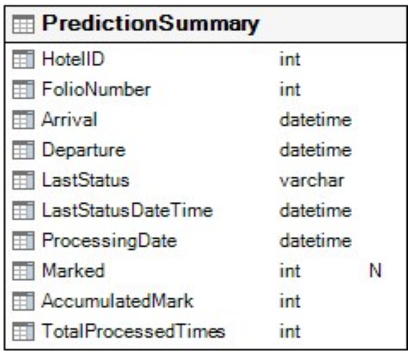


Table: BCPrototype.dbo.PredictionSummary

Description: Booking's prediction statistics per processing date

Name	Description	Data type	Max length	Nullable
HotelID	Hotel ID	int	4	
FolioNumber	Booking ID	int	4	
Arrival	Arrival date	datetime	8	
Departure	Departure date	datetime	8	
LastStatus	Booking last known status	varchar	1	
LastStatusDateTime	Date booking status was lastly update	datetime	8	
ProcessingDate	Processing date	datetime	8	
Marked	Indication of booking outcome prediction for the processing date (0: not-canceled; 1: canceled;)	int	4	✓
AccumulatedMark	Number of times the booking was marked as likely to cancel at the processing date	int	4	
TotalProcessedTimes	Total number of times the booking was processed until the processing date	int	4	

Data model



PredictionSummary

HotelID	int	
FolioNumber	int	
Arrival	datetime	
Departure	datetime	
LastStatus	varchar	
LastStatusDateTime	datetime	
ProcessingDate	datetime	
Marked	int	N
AccumulatedMark	int	
TotalProcessedTimes	int	

Table: BCPrototype.dbo.AspNetUserRoles

Description: List of roles each user belongs to

Name	Description	Data type	Max length	Nullable
UserId	User ID	nvarchar	128	
RoleId	Role ID	nvarchar	128	

Data model

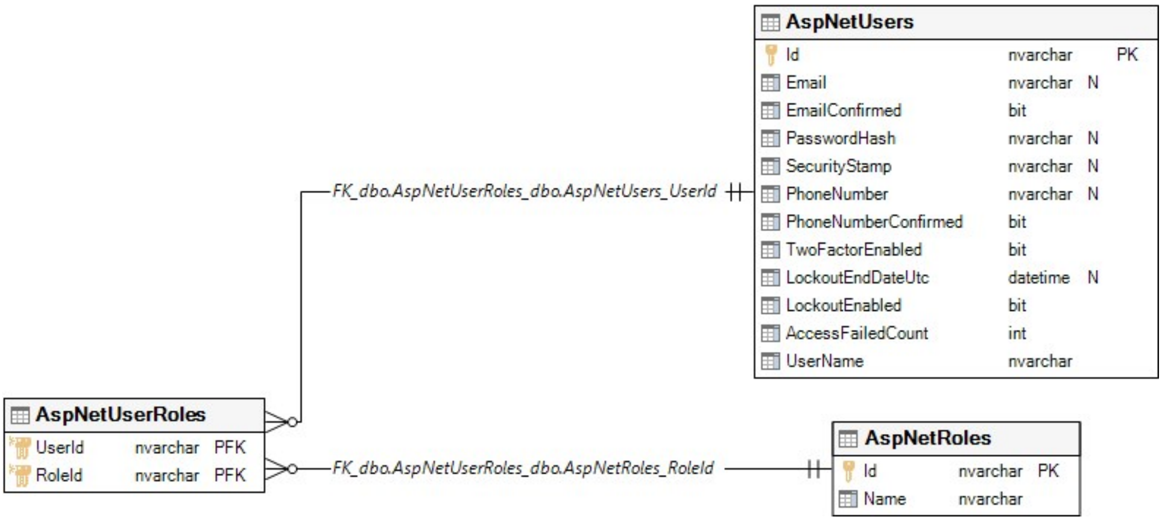


Table: BCPrototype.dbo.AspNetUsers

Description: List of system users

Name	Description	Data type	Max length	Nullable
Id	Internal ID	nvarchar	128	
Email	Email address	nvarchar	256	✓
EmailConfirmed	Indication if email is configured	bit	1	
PasswordHash	Password hash	nvarchar	1073741823	✓
SecurityStamp	Stamp to confirm data is not tampered	nvarchar	1073741823	✓
PhoneNumber	Phone number	nvarchar	1073741823	✓
PhoneNumberConfirmed	Indication if phone number is confirmed	bit	1	
TwoFactorEnabled	Indication if two factor authentication is enabled	bit	1	
LockoutEndDateUtc	Date and time when user was locked out (UTC format)	datetime	8	✓
LockoutEnabled	Indication if locked out mechanism is enabled	bit	1	
AccessFailedCount	Number of consecutive failed login attempts	int	4	
UserName	User name	nvarchar	256	

Data model

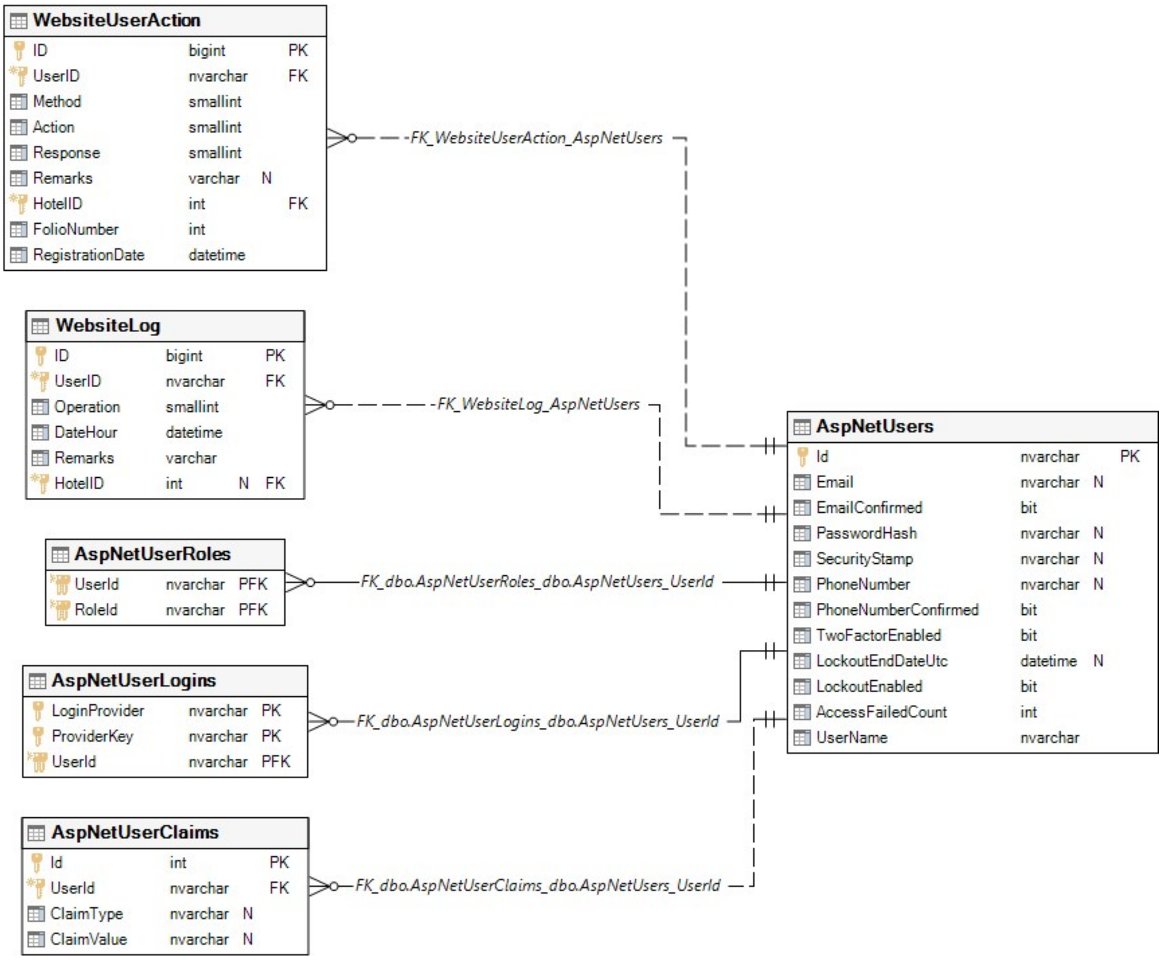


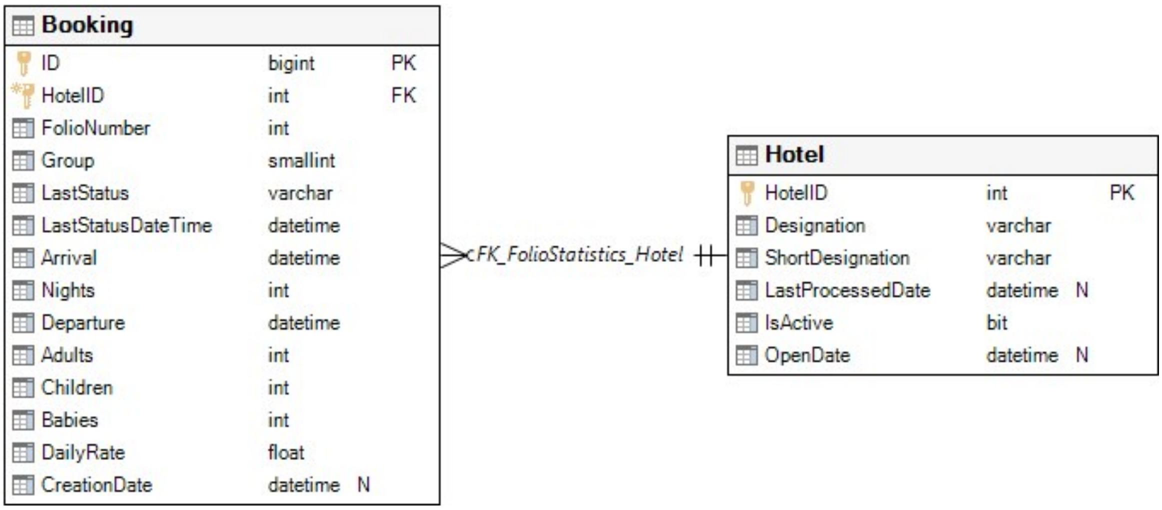
Table: BCPrototype.dbo.Booking

Description: Bookings details

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
HotelID	Hotel ID	int	4	
FolioNumber	PMS booking ID	int	4	
Group	A/B test group was assined to (0: Control ;1: Verification;)	smallint	2	
LastStatus	Booking status (C: Confirmed; G: Guarantee; N: No-show; A: Canceled; R: Checked-in; O: Check-out)	varchar	1	
LastStatusDateTime	Date and time the current status was assined to the booking	datetime	8	
Arrival	Arrival date	datetime	8	
Nights	Number of staying nights	int	4	
Departure	Departure date	datetime	8	
Adults	Number of adults	int	4	
Children	Number of children	int	4	

Name	Description	Data type	Max length	Nullable
Babies	Number of babies	int	4	
DailyRate	Average daily rate	float	8	
CreationDate	Date and time of booking creation	datetime	8	✓

Data model



Booking		
ID	bigint	PK
HotelID	int	FK
FolioNumber	int	
Group	smallint	
LastStatus	varchar	
LastStatusDateTime	datetime	
Arrival	datetime	
Nights	int	
Departure	datetime	
Adults	int	
Children	int	
Babies	int	
DailyRate	float	
CreationDate	datetime	N

Hotel		
HotelID	int	PK
Designation	varchar	
ShortDesignation	varchar	
LastProcessedDate	datetime	N
IsActive	bit	
OpenDate	datetime	N

Table: BCPrototype.dbo.ExecutionLog

Description: Log of models' daily processing

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ProcessingDate	Processing date	datetime	8	
HotelID	Hotel ID	int	4	
Start	Processing start date and time	datetime	8	
End	Processing end date and time	datetime	8	
AppliedModelID	ID of model which was applied	bigint	8	✓
NewModelID	ID of model that was developed on day of processing	bigint	8	✓
PreviousDayCurrentModelID	ID of model that was applied on the previous day	bigint	8	✓

Data model

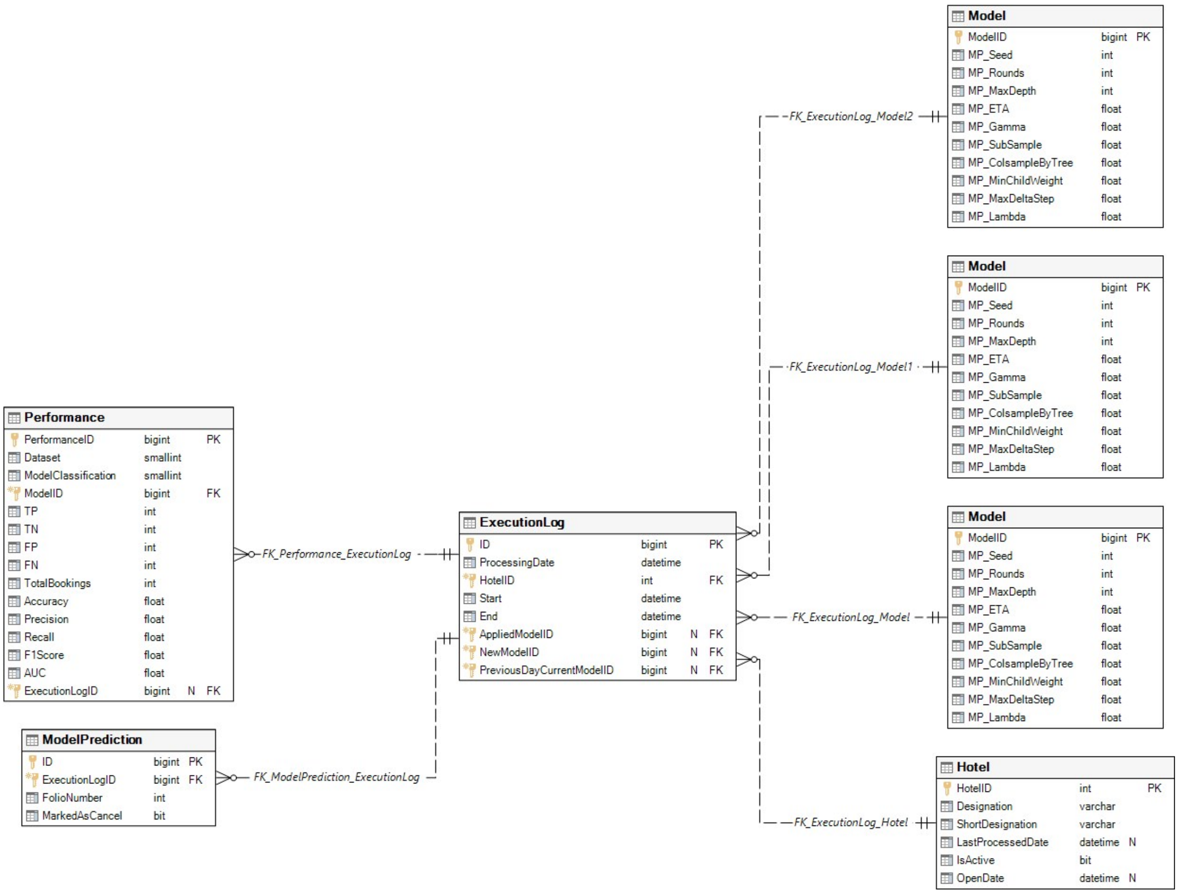
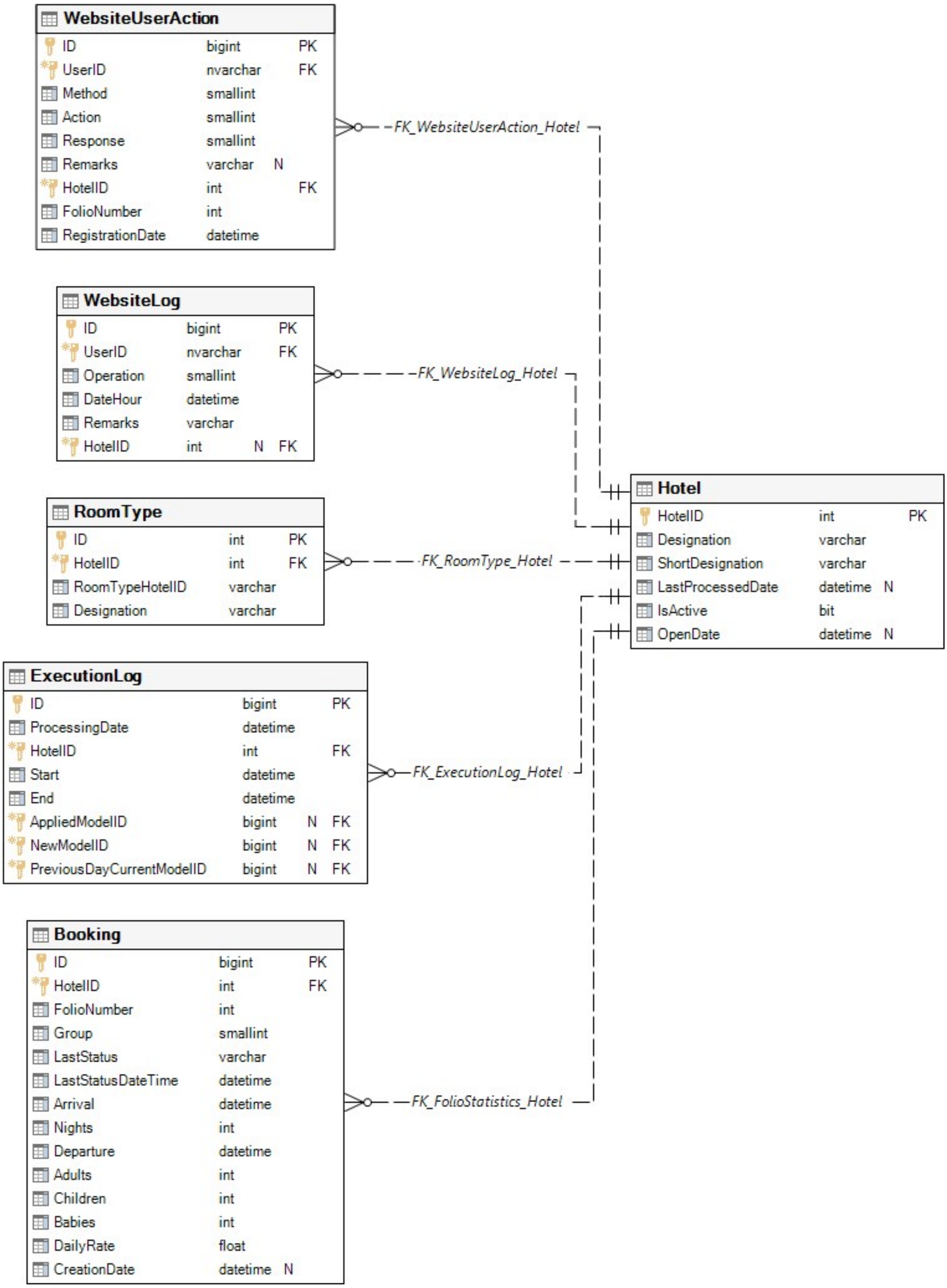


Table: BCPrototype.dbo.Hotel

Description: List of hotels

Name	Description	Data type	Max length	Nullable
HotelID	Internal ID	int	4	
Designation	Hotel name	varchar	50	
ShortDesignation	Hotel short name	varchar	5	
LastProcessedDate	Last date model was processed for the current hotel	datetime	8	✓
IsActive	Indication if processing for the current hotel should be made	bit	1	
OpenDate	Date when hotel started operating (for analytics operations)	datetime	8	✓

Data model



WebsiteUserAction

ID	bigint		PK
UserID	nvarchar		FK
Method	smallint		
Action	smallint		
Response	smallint		
Remarks	varchar	N	
HotelID	int		FK
FolioNumber	int		
RegistrationDate	datetime		

WebsiteLog

ID	bigint		PK
UserID	nvarchar		FK
Operation	smallint		
DateHour	datetime		
Remarks	varchar		
HotelID	int	N	FK

RoomType

ID	int		PK
HotelID	int		FK
RoomTypeHotelID	varchar		
Designation	varchar		

ExecutionLog

ID	bigint		PK
ProcessingDate	datetime		
HotelID	int		FK
Start	datetime		
End	datetime		
AppliedModelID	bigint	N	FK
NewModelID	bigint	N	FK
PreviousDayCurrentModelID	bigint	N	FK

Booking

ID	bigint		PK
HotelID	int		FK
FolioNumber	int		
Group	smallint		
LastStatus	varchar		
LastStatusDateTime	datetime		
Arrival	datetime		
Nights	int		
Departure	datetime		
Adults	int		
Children	int		
Babies	int		
DailyRate	float		
CreationDate	datetime	N	

Hotel

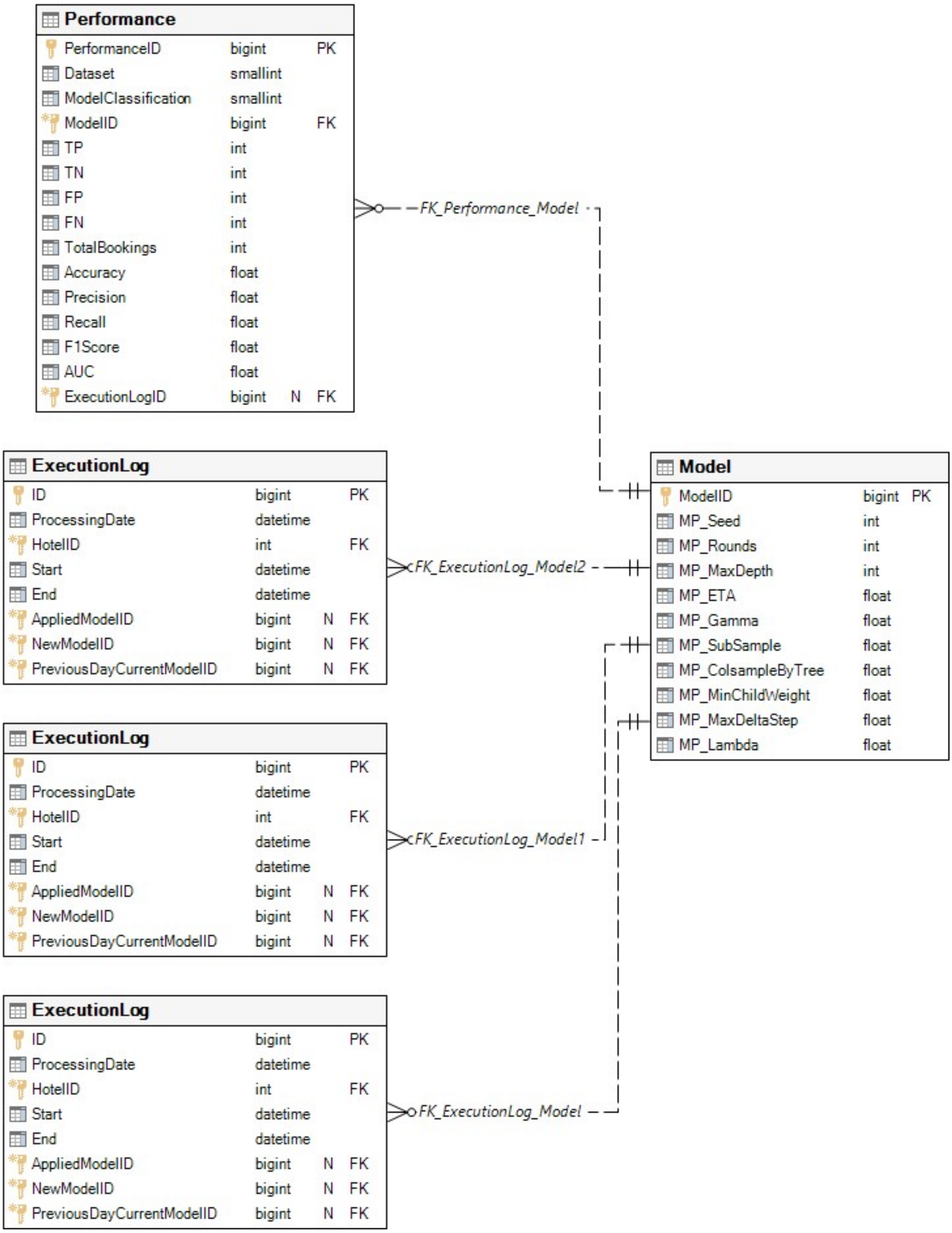
HotelID	int		PK
Designation	varchar		
ShortDesignation	varchar		
LastProcessedDate	datetime	N	
IsActive	bit		
OpenDate	datetime	N	

Table: BCPrototype.dbo.Model

Description: Built models parameters

Name	Description	Data type	Max length	Nullable
ModelID	Internal ID	bigint	8	
MP_Seed	Seed to use on the creation of random numbers	int	4	
MP_Rounds	Number of XGBoost rounds	int	4	
MP_MaxDepth	Trees maximum depth	int	4	
MP_ETA	XGBoost learning rate	float	8	
MP_Gamma	XGBoost gamma	float	8	
MP_SubSample	Sub sample of rows to include in each tree	float	8	
MP_ColsampleByTree	Sub sample of features to include in each tree	float	8	
MP_MinChildWeight	XGBoost minimum chield weight	float	8	
MP_MaxDeltaStep	XGBoost delta step	float	8	
MP_Lambda	Lambda regularization	float	8	

Data model



Performance

PerformanceID	bigint		PK
Dataset	smallint		
ModelClassification	smallint		
ModelID	bigint		FK
TP	int		
TN	int		
FP	int		
FN	int		
TotalBookings	int		
Accuracy	float		
Precision	float		
Recall	float		
F1Score	float		
AUC	float		
ExecutionLogID	bigint	N	FK

ExecutionLog

ID	bigint		PK
ProcessingDate	datetime		
HotelID	int		FK
Start	datetime		
End	datetime		
AppliedModelID	bigint	N	FK
NewModelID	bigint	N	FK
PreviousDayCurrentModelID	bigint	N	FK

Model

ModelID	bigint	PK
MP_Seed	int	
MP_Rounds	int	
MP_MaxDepth	int	
MP_ETA	float	
MP_Gamma	float	
MP_SubSample	float	
MP_ColsampleByTree	float	
MP_MinChildWeight	float	
MP_MaxDeltaStep	float	
MP_Lambda	float	

ExecutionLog

ID	bigint		PK
ProcessingDate	datetime		
HotelID	int		FK
Start	datetime		
End	datetime		
AppliedModelID	bigint	N	FK
NewModelID	bigint	N	FK
PreviousDayCurrentModelID	bigint	N	FK

ExecutionLog

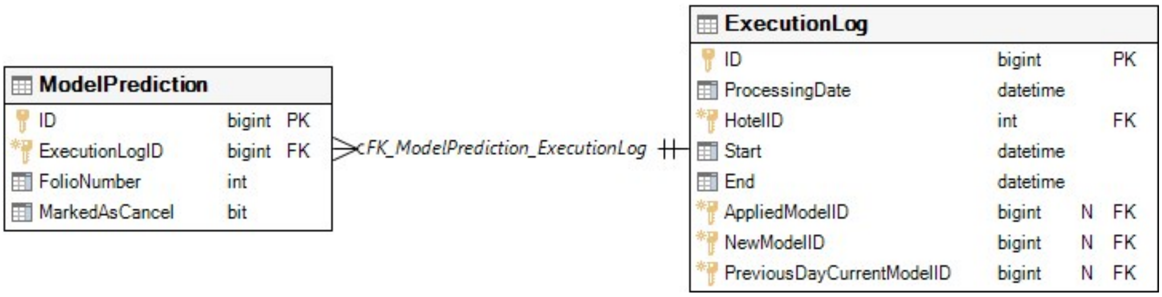
ID	bigint		PK
ProcessingDate	datetime		
HotelID	int		FK
Start	datetime		
End	datetime		
AppliedModelID	bigint	N	FK
NewModelID	bigint	N	FK
PreviousDayCurrentModelID	bigint	N	FK

Table: BCPrototype.dbo.ModelPrediction

Description: Models' predictions per execution and booking

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
ExecutionLogID	Execution log ID	bigint	8	
FolioNumber	Booking ID	int	4	
MarkedAsCancel	Prediction outcome (0: not-canceled; 1: canceled;)	bit	1	

Data model



ModelPrediction

ID	bigint	PK
ExecutionLogID	bigint	FK
FolioNumber	int	
MarkedAsCancel	bit	

ExecutionLog

ID	bigint		PK
ProcessingDate	datetime		
HotelID	int		FK
Start	datetime		
End	datetime		
AppliedModelID	bigint	N	FK
NewModelID	bigint	N	FK
PreviousDayCurrentModelID	bigint	N	FK

Table: BCPrototype.dbo.Performance

Description: Models' execution statistics

Name	Description	Data type	Max length	Nullable
PerformanceID	Internal ID	bigint	8	
Dataset	Dataset (0: train; 1: test;)	smallint	2	
ModelClassification	Model classification (0: current model; 1: new model	smallint	2	
ModelID	Model ID	bigint	8	
TP	True positives	int	4	
TN	True negatives	int	4	
FP	False positives	int	4	
FN	False negatives	int	4	
TotalBookings	Total bookings processed	int	4	
Accuracy	Accuracy	float	8	
Precision	Precision	float	8	
Recall	Recall	float	8	
F1Score	F1Score	float	8	
AUC	AUC	float	8	
ExecutionLogID	Execution log ID it belongs to	bigint	8	✓

Data model

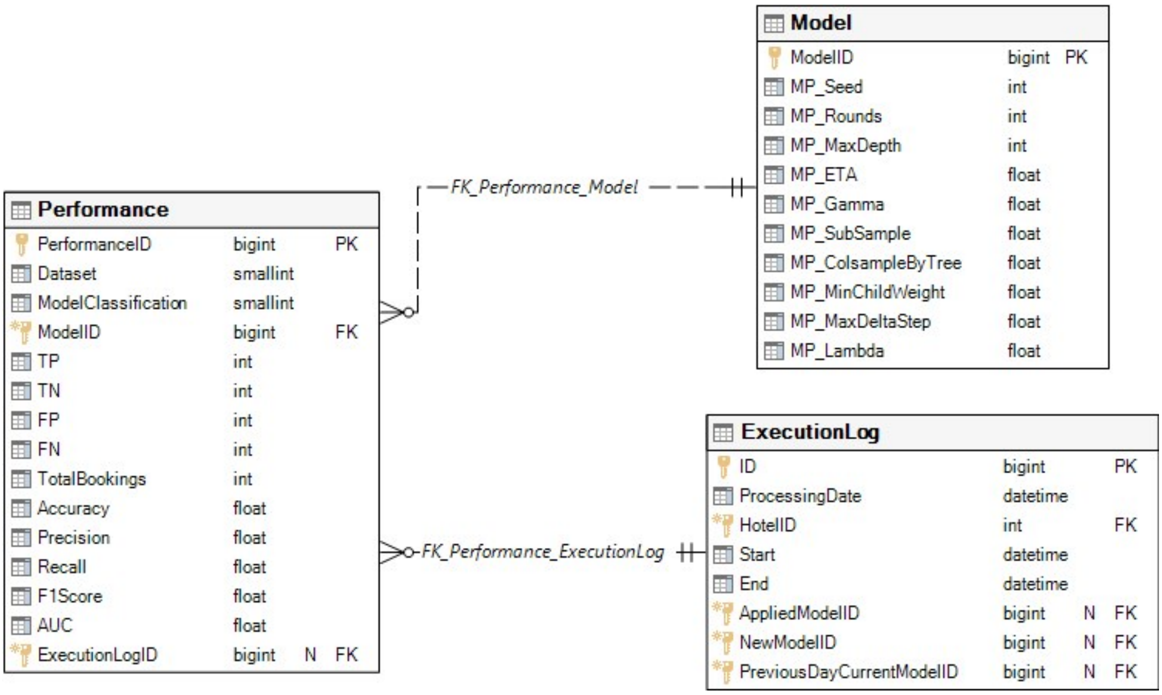


Table: BCPrototype.dbo.RoomType

Description: Hotels' room types

Name	Description	Data type	Max length	Nullable
ID	Internal ID	int	4	
HotelID	Hotel ID	int	4	
RoomTypeHotelID	Room type ID	varchar	1	
Designation	Designation	varchar	5	

Data model

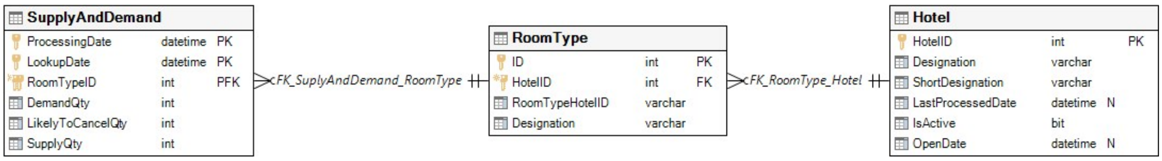


Table: BCPrototype.dbo.SupplyAndDemand

Description: Totals of rooms available for sale in inventory and rooms sold per day

Name	Description	Data type	Max length	Nullable
ProcessingDate	Processind date	datetime	8	
LookupDate	Lookup date (for every date in the future where the is a booking)	datetime	8	

Name	Description	Data type	Max length	Nullable
RoomTypeID	Room type ID	int	4	
DemandQty	Number of bookings for the lookup day, at the processing date	int	4	
LikelyToCancelQty	Number of bookings predicted to cancel at the lookup date, at the processing date	int	4	
SupplyQty	Number of rooms availabe (not out-of-order and not out-of-service)	int	4	

Data model

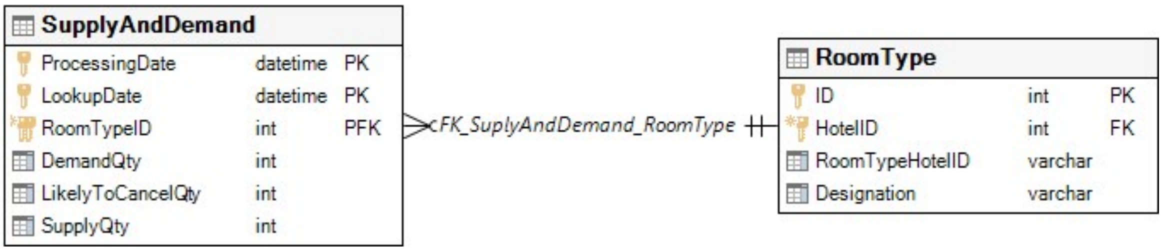


Table: BCPrototype.dbo.WebsiteLog

Description: Log of operations conducted on the website

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
UserID	User ID	nvarchar	128	
Operation	Operation description	smallint	2	
DateHour	Date and time	datetime	8	
Remarks	Remarks to the operation	varchar	100	
HotelID	Hotel ID	int	4	✓

Data model

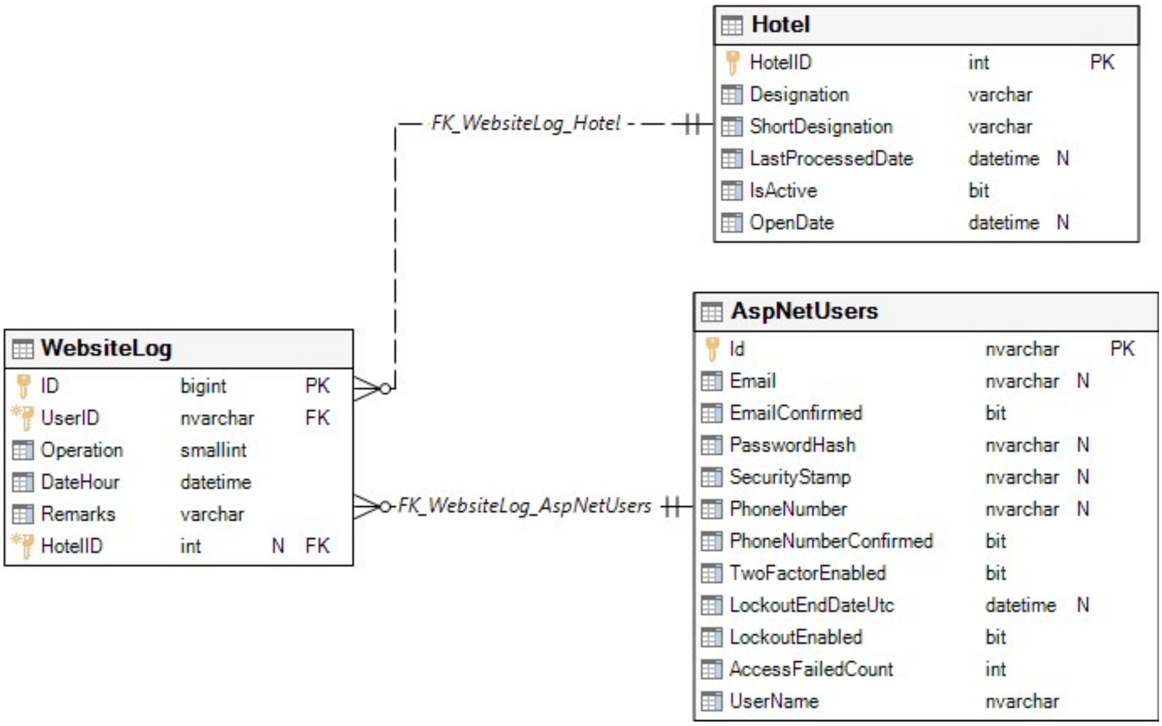
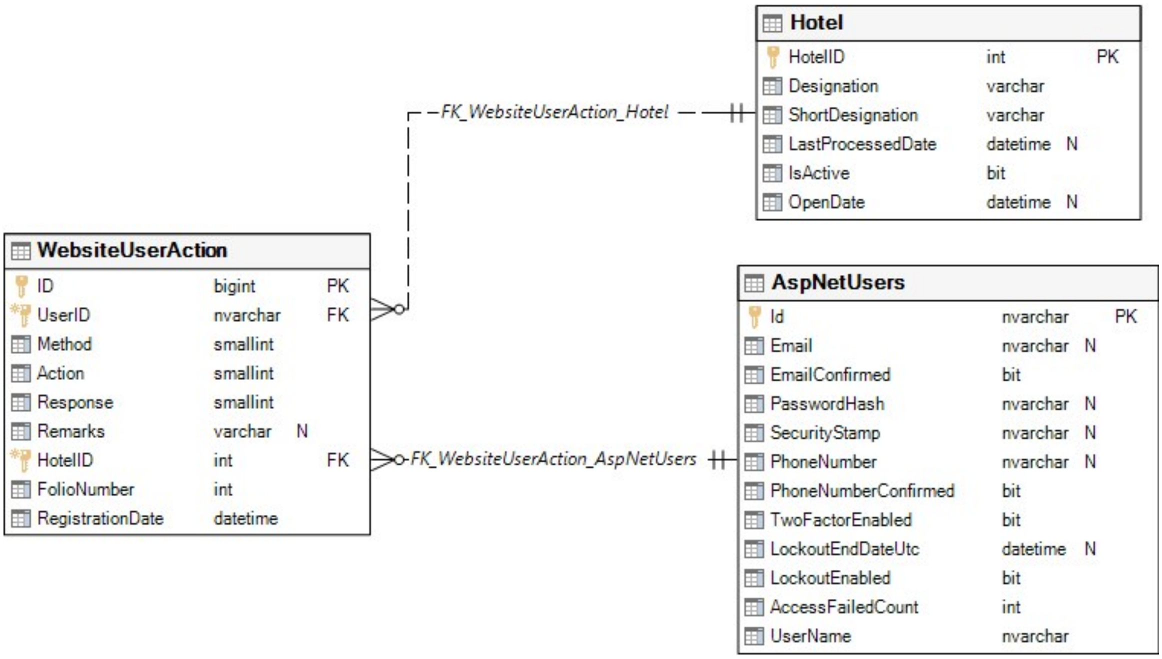


Table: BCPrototype.dbo.WebsiteUserAction

Description: Log of customers contacts to prevent cancellation

Name	Description	Data type	Max length	Nullable
ID	Internal ID	bigint	8	
UserID	User ID	nvarchar	128	
Method	Method the user employed to contact the customer (0: phone; 1: email; 2: other;)	smallint	2	
Action	Type of action made (0: discounts; 1: services; 2: upgrade; 3: other;)	smallint	2	
Response	Type of response obtained from customer (0: accepted; 1: canceled; 2: other;)	smallint	2	
Remarks	Additional remarks	varchar	500	✓
HotelID	Hotel ID	int	4	
FolioNumber	Booking ID	int	4	
RegistrationDate	Date and time when contact was logged in the website	datetime	8	

Data model



Hotel		
HotelID	int	PK
Designation	varchar	
ShortDesignation	varchar	
LastProcessedDate	datetime N	
IsActive	bit	
OpenDate	datetime N	

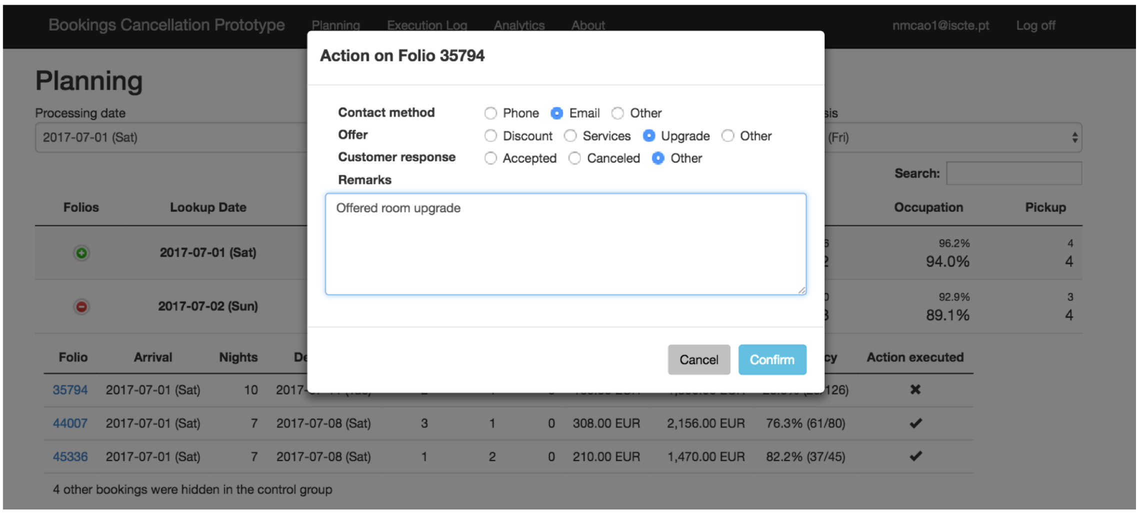
WebsiteUserAction			
ID	bigint		PK
UserID	nvarchar		FK
Method	smallint		
Action	smallint		
Response	smallint		
Remarks	varchar	N	
HotelID	int		FK
FolioNumber	int		
RegistrationDate	datetime		

AspNetUsers			
Id	nvarchar		PK
Email	nvarchar	N	
EmailConfirmed	bit		
PasswordHash	nvarchar	N	
SecurityStamp	nvarchar	N	
PhoneNumber	nvarchar	N	
PhoneNumberConfirmed	bit		
TwoFactorEnabled	bit		
LockoutEndDateUtc	datetime	N	
LockoutEnabled	bit		
AccessFailedCount	int		
UserName	nvarchar		

APPENDIX G – PROTOTYPE WEB PAGES EXAMPLES

This appendix shows examples of the pages of the prototype not shown in Chapter 5.

Example of a page users accessed to report researchers the details of a booking that was predicted as likely to cancel and that the users contacted in an effort to try to prevent its cancellation.



Planning

Processing date: 2017-07-01 (Sat)

Action on Folio 35794

Contact method ○ Phone ● Email ○ Other
Offer ○ Discount ○ Services ● Upgrade ○ Other
Customer response ○ Accepted ○ Canceled ● Other
Remarks

Offered room upgrade

(Fri)

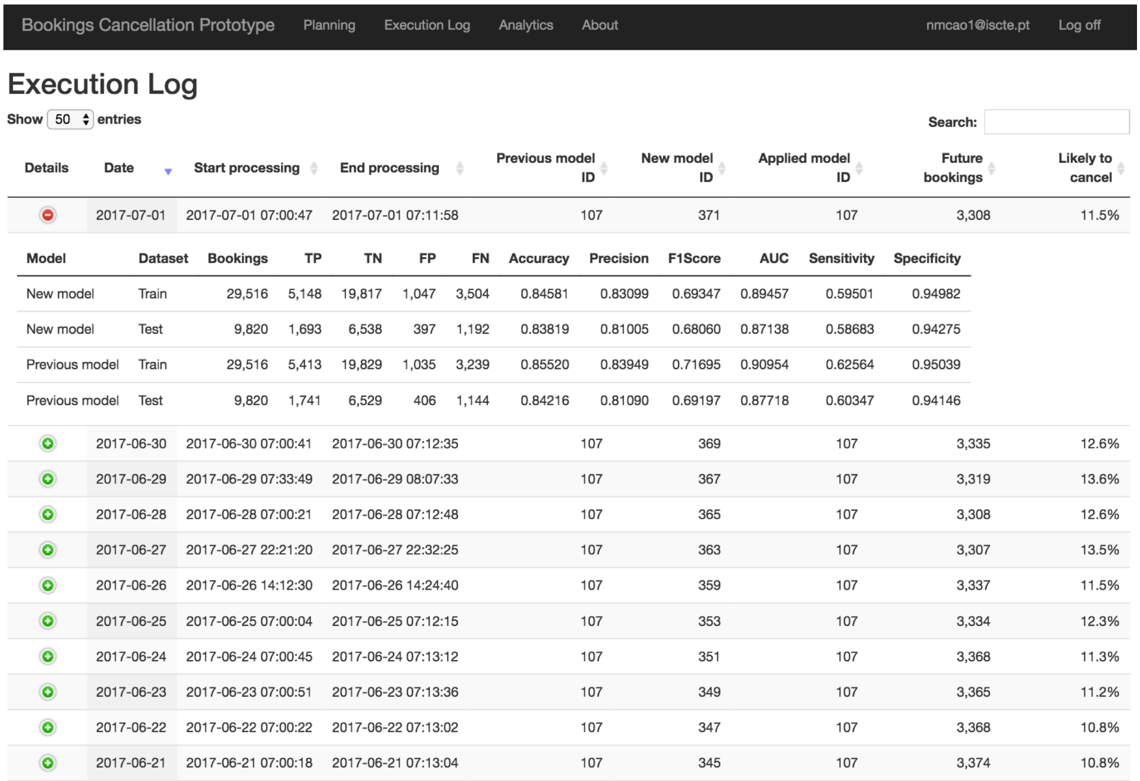
Search:

Folios	Lookup Date	Occupation	Pickup
⊕	2017-07-01 (Sat)	96.2% 94.0%	4 4
⊖	2017-07-02 (Sun)	92.9% 89.1%	3 4

Folio	Arrival	Nights	De						cy	Action executed
35794	2017-07-01 (Sat)	10	2017-						126)	✖
44007	2017-07-01 (Sat)	7	2017-07-08 (Sat)	3	1	0	308.00 EUR	2,156.00 EUR	76.3% (61/80)	✓
45336	2017-07-01 (Sat)	7	2017-07-08 (Sat)	1	2	0	210.00 EUR	1,470.00 EUR	82.2% (37/45)	✓

4 other bookings were hidden in the control group

Example of “Execution log” page. In this page, researchers and users could consult the system performance metrics, summary statistics on predictions, and details on the model applied.



Execution Log

Show 50 entries

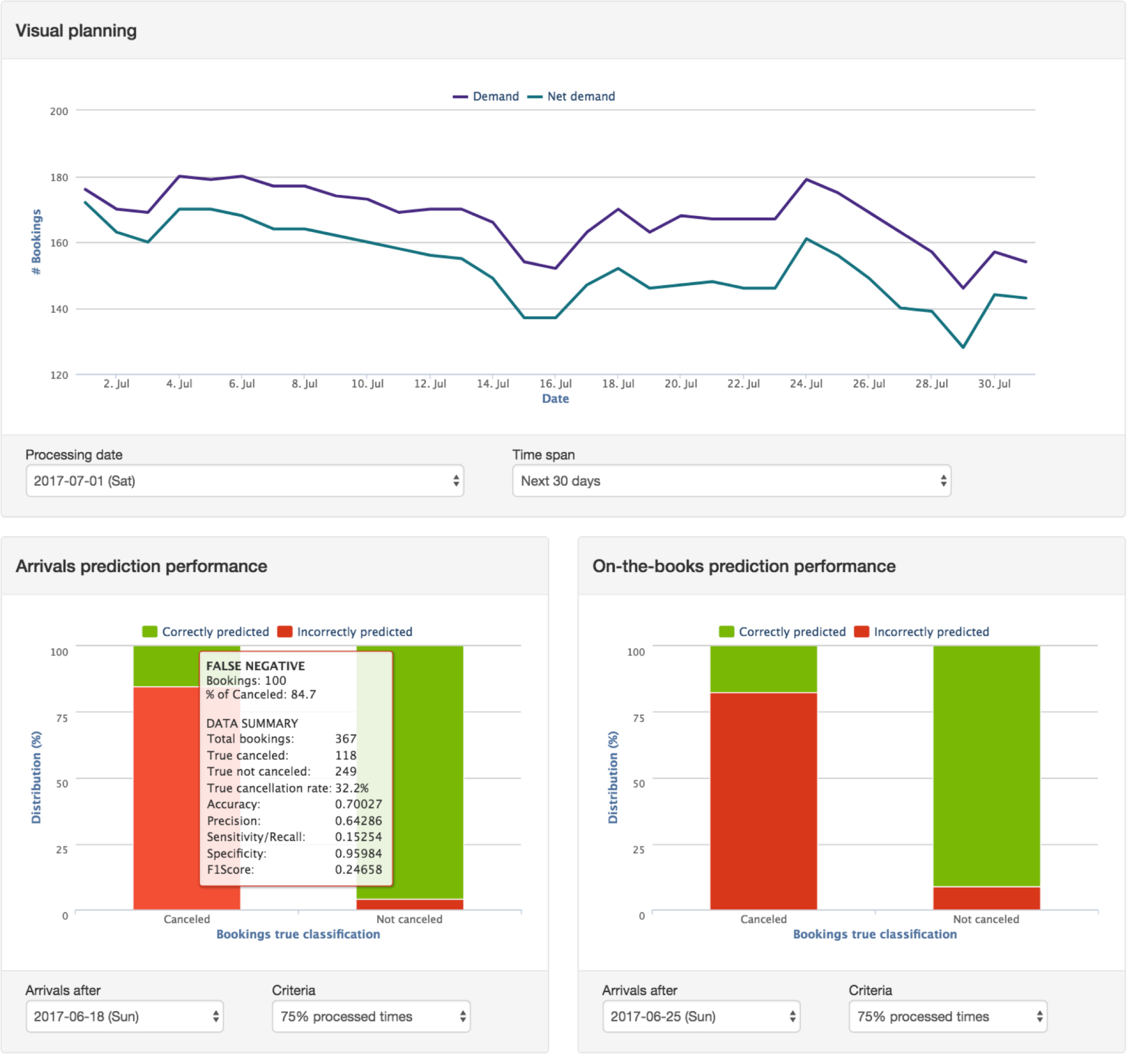
Search:

Details	Date	Start processing	End processing	Previous model ID	New model ID	Applied model ID	Future bookings	Likely to cancel
⊖	2017-07-01	2017-07-01 07:00:47	2017-07-01 07:11:58	107	371	107	3,308	11.5%

Model	Dataset	Bookings	TP	TN	FP	FN	Accuracy	Precision	F1Score	AUC	Sensitivity	Specificity
New model	Train	29,516	5,148	19,817	1,047	3,504	0.84581	0.83099	0.69347	0.89457	0.59501	0.94982
New model	Test	9,820	1,693	6,538	397	1,192	0.83819	0.81005	0.68060	0.87138	0.58683	0.94275
Previous model	Train	29,516	5,413	19,829	1,035	3,239	0.85520	0.83949	0.71695	0.90954	0.62564	0.95039
Previous model	Test	9,820	1,741	6,529	406	1,144	0.84216	0.81090	0.69197	0.87718	0.60347	0.94146

Details	Date	Start processing	End processing	Previous model ID	New model ID	Applied model ID	Future bookings	Likely to cancel
⊕	2017-06-30	2017-06-30 07:00:41	2017-06-30 07:12:35	107	369	107	3,335	12.6%
⊕	2017-06-29	2017-06-29 07:33:49	2017-06-29 08:07:33	107	367	107	3,319	13.6%
⊕	2017-06-28	2017-06-28 07:00:21	2017-06-28 07:12:48	107	365	107	3,308	12.6%
⊕	2017-06-27	2017-06-27 22:21:20	2017-06-27 22:32:25	107	363	107	3,307	13.5%
⊕	2017-06-26	2017-06-26 14:12:30	2017-06-26 14:24:40	107	359	107	3,337	11.5%
⊕	2017-06-25	2017-06-25 07:00:04	2017-06-25 07:12:15	107	353	107	3,334	12.3%
⊕	2017-06-24	2017-06-24 07:00:45	2017-06-24 07:13:12	107	351	107	3,368	11.3%
⊕	2017-06-23	2017-06-23 07:00:51	2017-06-23 07:13:36	107	349	107	3,365	11.2%
⊕	2017-06-22	2017-06-22 07:00:22	2017-06-22 07:13:02	107	347	107	3,368	10.8%
⊕	2017-06-21	2017-06-21 07:00:18	2017-06-21 07:13:04	107	345	107	3,374	10.8%

Example of the “Analytics” page. In this page, researcher and users could visualize and explore, analytically, performance metrics and predictions statistics.



Visual planning

Processing date: 2017-07-01 (Sat)

Time span: Next 30 days

Arrivals prediction performance

FALSE NEGATIVE
Bookings: 100
% of Canceled: 84.7

DATA SUMMARY

Total bookings:	367
True canceled:	118
True not canceled:	249
True cancellation rate:	32.2%
Accuracy:	0.70027
Precision:	0.64286
Sensitivity/Recall:	0.15254
Specificity:	0.95984
F1Score:	0.24658

Arrivals after: 2017-06-18 (Sun)

Criteria: 75% processed times

On-the-books prediction performance

Arrivals after: 2017-06-25 (Sun)

Criteria: 75% processed times