Department of Marketing, Operation and Management

Stripping customers' feedback on hotels evaluation through data mining

Joana Pinto Coelho

A Dissertation presented in partial fulfillment of the Requirements for the Degree of Master in Hospitality and Tourism Management

Supervisor: Paulo Rita, PhD, Full Professor, ISCTE-University Institute of Lisbon

Co-supervisor: Sérgio Moro, PhD, Assistant Professor, ISCTE-University Institute of Lisbon

[September, 2016]

Resumo

Com a constante evolução tecnológica e a consequente afluência de partilha de informação entre os consumidores, as plataformas online, como é o caso do TripAdvisor, começaram a ser usadas para análise, principalmente na indústria hoteleira. Estas plataformas permitem aos clientes a partilha de opiniões e a respectiva atribuição de uma avaliação quantitativa aos hotéis visitados. Os estudos publicados têm-se focado, fundamentalmente, na análise dos comentários; contudo, estudos relacionados com a avaliação quantitativa são mais escassos.

Este estudo foi desenvolvido através de técnicas de *data mining* por forma a modelar a pontuação atribuída no TripAdvisor. Foram recolhidos dois comentários por cada mês do ano de 2015 referentes a 21 hotéis localizados na avenida mais emblemática de Las Vegas, a Strip, num total de 504 comentários. A localização foi seleccionada por ser um destino de elevado impato turístico já que a cidade persiste devido à hotelaria e aos casinos. Foram seleccionadas 19 variáveis que representam o utilizador, o hotel e as suas características para alimentarem uma máquina de vectores de suporte objectivando a modelação da avaliação quantitativa para extração de conhecimento. Os resultados atestaram a utilidade do modelo na sua capacidade preditiva. Após esta validação foi aplicada uma análise de sensibilidade ao modelo para compreender a relevância das variáveis.

Os resultados revelaram que as variáveis diretamente relacionadas com o utilizador e a sua experiência na utilização do TripAdvisor têm maior influência na atribuição das pontuações, comparativamente com as variáveis relacionadas com o hotel.

Sistema de classificação (JEL)

Z32 Tourism and Development

M31 Marketing

Abstract

The emergence of online reviews' platforms such as TripAdvisor provided tools for tourists to write their opinions and rate hotels with a quantitative score. While numerous studies are found based on textual comments of users, research on the score is rather scarce.

This study presents a data mining approach for modeling TripAdvisor score using 504 reviews published in 2015 for the 21 hotels located in the Strip, Las Vegas. Nineteen features characterizing the reviews, hotels and the users were prepared and used for feeding a support vector machine for modeling the score. The results achieved reveal the model is a good approximation for predicting the score. Therefore, a sensitivity analysis was applied over the model for extracting useful knowledge translated into features' relevance for the score. The findings unveiled user features related to TripAdvisor membership experience play a key role in influencing the scores granted, clearly surpassing hotel features.

Keywords

Customer feedback; customer reviews; online reviews; knowledge extraction; data mining; modeling; sensitivity analysis; Las Vegas.

JEL classification system:

Z32 Tourism and Development

M31 Marketing

Contents

Tables Index

Figures Index

1. Introduction

The Online Travel Agencies (OTA) are now the most used tool of travel booking, both for the means of transport and accommodation (Mauri & Minazzi, 2013) and, consequently, online reviews have been exponentially increasing its use and impact in the hospitality industry over the last years, due to the social media and technological evolution. In fact, according to Vermeulen and Seegers (2009), online reviews contribute with $10 billion in online travel purchases, as over 60% of the potential hotel customers search for online feedback before travelling and base their purchase decisions on it (Mauri & Minazzi, 2013). Therefore, electronic word-of-mouth (eWOM), which according to Henning-Thurau et al. (2004) is defined as "any positive or negative statement made by potential, actual or former customers about a product or company, which is made available to a multitude of people and institutions via the internet", has become a huge aspect when travelling, since nowadays every consumer has access to the internet and can easily express either positive or negative feedback. Most importantly, it is an online tool to be used when others seek for advice as part of the decision-making process, such as where to stay, especially in hospitality industry, as consumers are purchasing an experience and cannot predict its evaluation (Sparks & Browning, 2011). Therefore, holidays can be considered as a high risk and involvement purchase, due to its usual personal importance and also high value of money (Papathanassis & Knolle, 2009). In every industry, service quality is a determinant of the customer's perceptions and their feedback. The ideal would be that the target's expectations meet the perceptions, which will directly influence a positive word of mouth, contributing for a development of reputation and trust (Corbitt & Thanasankit, 2003). Hence, research contributions that unveil and provide in-depth understanding on the features that have the most impact on customer feedback are valuable for sustainable decision making.

Previous studies have been conducted by various researchers in order to understand and explain the influence and impact of online reviews in the hospitality industry. One of the most common methods used include the analysis of variance (ANOVA) technique, which is offered in many data analysis' solutions such as the IBM SPSS software. For example, Vermeulen and Seegers (2009) adopted the ANOVA for testing whether or not the user-generated online reviews influence the consumer choice. Additionally, Sparks and Browning (2011) went further on their

research and studied the fact that a consumer generated quantitative rating could be associated together with the actual written review. In a more recent data-driven study, it has been showed through regression models that the financial benefits of an online review from TripAdvisor conceal intrinsic value to the hospitality industry (Neirotti et al., 2016). Nevertheless, the majority of previous recent studies are focused on the impact of the text review itself, applying text mining techniques, which aim to extract meaningful knowledge from a variety of textual data and find relationships and patterns within such unstructured information (He et al., 2013). Different studies are aligned through similar conclusions regarding the fact that text mining applications to social media data (i.e. any online platform where customers can exchange information) can provide significant insights on the human behavior and interaction (e.g., He et al., 2013). However, while several studies are known using data mining for sentiment classification and opinion mining (e.g., Schuckert et al., 2015), none was found up to the present adopting a quantitative approach on modeling tourists' reviews through advanced data mining techniques for extracting the influence of hotels' and users' features on the score provided by users. Hence, the present study aims at filling such gap by focusing on online reviews' quantitative features such as number of stars of the hotel and number of helpful votes the user has received for building a predictive model of the tourists' score on the hotels. The knowledge built upon such model may help to shed some light on what drives the rating of a hotel, potentiating meaningful information to support managerial decisions.

Las Vegas, the so called city of sin, was the elected location for the present research. This hospitality city, born eighty years ago over a desert where hotels started to be built and forming one of the most entertaining cities in the world, is driven by the tourism and the gambling pleasure (Rowley, 2015). According to the US Bureau of Economic Analysis, due to its continuously growing and transforming, Las Vegas has the fastest growth rate in the United States and its personal income average ranks 34th in the country. Regarding previous studies conducted about and within Las Vegas mainly in the Strip, the most popular avenue of the city and with the largest supply of hotel rooms, Ro et al. (2013) discussed the affective image of the major hotel's positioning, while the city's success as a gaming destination due to the government and private institutions was proposed and analyzed by Lee (2015).

Given the interest Las Vegas hospitality rises, the present research started by collecting all the features available on TripAdvisor's webpages from several online reviews published during 2015 and targeting hotels located in the Strip. Thereafter, such dataset was modeled according to the score given by users through a support vector machine algorithm, with the resulting model being evaluated in terms of its predictive performance to assess the most likely rank for each review. Finally, the model was opened using a sensitivity analysis method for extracting useful knowledge in terms of which of the input features known prior to the review influenced most the outcome score. Such approach is an attempt to answer the following raised questions: Can the score of an online hospitality review be predicted using as input only quantitative data? What are the features that influence most the review scores in hospitality? How does each of those features affect the score and can this knowledge be useful for hotel managers?

Concluding, the main goals and contributions of this study are as follows:

- Creating a model that predicts the review score based on quantitative features of the user/reviewer and the hotel, as well as the period of time of the specific stay;
- Contributing to research on customers' feedback and online reviews by providing a novel approach on the data used, the quantitative features, as opposed to the most common analyses of the reviews' text itself;
- Understanding how users are inherently influenced by hotels' features when submitting numerical scores besides text comments on online platforms, such as TripAdvisor.

The next section will describe the background concepts, such as the history and evolution of online reviews, as well as the methods for knowledge extraction from data, its dimensions and its use in the industry. The following section will discuss the materials (e.g. input dataset) and procedures that were applied in the experiment. Then, the results are shown and a critical discussion takes place on the findings section. Finally, the main conclusions of this research are drawn.

2. Theory

2.1. Online reviews

In 2004, Tim O'Reilly coined the term Web 2.0 as the network connecting all devices to which individual users contribute largely by sharing their experiences in numerous ways, therefore becoming one of the most relevant sources of the internet through the so called user-generated contents (O'Reilly & Battelle, 2009). Such internet evolution effectively became a global revolution, including the tourism and hospitality industry by adding new online sources of information to the existing hotel and tourism companies' websites, implying users are becoming key-players in influencing others through their online reviews (Papathanassis & Knolle, 2009).

Traditional websites have therefore evolved by increasing interactivity level to keep pace with Web 2.0 new demands. However, in this new information-driven era, specialized user-contents sites and applications such as wikis, forums, blogs, social networks and especially online reviews' sites for the case of tourism and hospitality have underpinned a new paradigm in which the user is at the center of the network, leading to a mutual exchange and sharing of values (Mazurek, 2009). Duan et al. (2008) particularly emphasize the impact of Web 2.0 in tourism and hotels by stating metaphorically on their research that online reviews are just like a huge "megaphone" on promoting product sales.

As mentioned above, the eWOM generated through online reviews can be translated in either a positive or negative sentiment regarding certain institutions. Although a positive sentiment can influence consumers and potential ones in their final decision, it is most likely that a negative one would have a major impact on such individuals, since consumers tend to share and complain more about their unpleasant experiences (Breazeale, 2009). Not only is this evolution important to consumers and to their experiences, but also it contributes to the industry institutions' reputation and creates an opportunity for the companies to assimilate the online contributions and suggestions in order to grow and improve, gaining advantages in the market positioning (Papathanassis & Knolle, 2009).

Several studies are found based on online reviews for tourism and hospitality, especially to analyze how exchanges of information influence directly the consumer choices regarding a certain hotel (e.g., Vermeulen & Seegers, 2009), with most of them concluding that an exposure to an online hotel positive review will increase the average probability of that consumer to book a room in the same hotel. Therefore, existing literature acknowledges the influence of online reviews over a wide spectrum of stakeholders, from users to business managers, as the study by Ye et al. (2009a) showed by evaluating its impact in terms of business performance. Moreover, it was also concluded by Sparks and Browning (2011) that the use of categories, while consulting OTA's, such as the most recent reviews or even the numerical rating given by the user to classify a certain hotel can be valuable indicators for assisting decision making. Another interesting point was mentioned by Gretzel and Yoo (2008), where the authors argued that 90% of the consumers using TripAdvisor consider that such reviews and comments are helpful in terms of alerting which places and services need to be avoided as well as, in another positive term, which products and destinations suit better their needs.

2.2. Data mining

According to Sharda et al. (2015), data mining is "the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large databases". Data mining usage virtually spreads across any field of research from where data analysis is in demand. For example, it is mostly used for companies in order to analyze customer data within the customer relationship management (CRM) structure (Ngai et al., 2009). Due to its nature originated in both statistical and machine learning fields, data mining focuses on the machine-driven model building instead of hypothesis testing supervised by a specialized researcher (Magnini et al., 2003). Furthermore, it was discussed by the same researchers that data mining techniques create patterns that can be used in order to strengthen the relationship between the hotel and the frequent consumers, predicting the potential value of each customer and avoiding the cost of attracting new ones. Also in hospitality, by clustering the customers (e.g., through traveler type) it is possible for the company to know

their target and therefore be more efficient in satisfying their needs. It is also an important tool for the marketing department, since with this information it is possible to previously create personalized advertisements or create direct-mail campaigns for instance (Magnini et al., 2003).

A data mining project usually consists in cycles of relevant consecutive stages such as data understanding, preparation, modeling and evaluation (Moro et al., 2014). A few methodologies have emerged for defining guidelines to conduct a data mining project, such as the CRISP-DM (Chapman et al., 2000). One of the most critical steps in data mining is data preparation for modeling, which includes feature selection and feature engineering, i.e., choosing the variables that best characterize the problem and, if needed, compute or obtain additional features (Domingos, 2012; Moro et al., 2016a).

Although text mining is one of the most common techniques when analyzing online reviews, as it establishes patterns that determine trends through textual comments (Lau et al., 2005), this study focused on assessing the patterns hidden in the quantitative fields from TripAdvisor, instead of the textual review itself. Thus, as the problem is to model the score (the outcome to predict) attributed by users through the remaining features (the inputs), it becomes a supervised learning problem. Therefore, for modeling, the support vector machine was chosen, as it is one of the most advanced supervised learning techniques, by transforming inputs into a high m-dimensional feature space, using a nonlinear mapping. Consequently, the algorithm will fit its way to the best linear separating hyper plane, connected through the distributed set of support vector points, which will determine the support vector in the feature space, thus providing an accurate performance (Moro et al., 2016b)

While the high level of accuracy of support vector machines makes of them attractive to use, the inherent complexity makes them unreadable by a human user, as opposed to regression or decision tree models (Cortez & Embrechts, 2013). For opening such types of “black-box” models, from which neural networks are also an example, a few techniques can be used. Hence, knowledge extraction from complex models can be achieved through rule extraction or sensitivity analysis (Moro et al., 2014). The latter applies changes in the inputs through their range of possible values and evaluates how it affects the predicted output value (Palmer et al., 2006). Cortez and Embrechts (2013) further developed the sensitivity analysis method by proposing a

data-based sensitivity analysis (DSA) that takes advantage of the data used for training the model to assess multiple variations of the input features, thus evaluating the influence each feature exerts on the remaining ones, besides the impact on the outcome feature. The DSA has been adopted with success for extracting knowledge from models in a wide variety of studies such as wine modeling (Cortez et al., 2009), jet grouting (Tinoco et al., 2011) and bank telemarketing (Moro et al., 2014), and it was therefore also chosen for the present study.

Considering the score available for users to rate hotels in TripAdvisor is an integer value between 1 and 5, with 1 representing the lowest and 5 the highest scores respectively, the problem becomes a regression problem (Sharda et al., 2015), where the model needs to fit the input data for modeling the numerical outcome. Thus, two according metrics were adopted for computing model accuracy: the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE). The MAE is the mean of all absolute differences between the real value and the one predicted by the model, therefore it is the deviation of the capacity that the model has of predicting the correct value. The MAPE metric is the mean of all absolute differences between the real value and the one predicted by the model divided by the real score, in order to extract a percentage regarding each deviation. Both metrics are described in detail by Hyndman and Koehler (2006). One of the disadvantages of MAPE is that it becomes undetermined for outcome values near zero. Nevertheless, such issue does not apply to the present study, since the outcome varies from 1 to 5.

2.3. Data mining in tourism and hospitality

A large amount of studies by different authors were conducted where data mining procedures were undertaken on tourism and hospitality data. Min et al. (2002) studied the application of data mining, more specifically using decision tree modeling in order to develop the profile of a certain group of customers within different hotels. In another article, data mining has also been studied regarding its importance and influence in a hotel’s marketing department and how it can help providing a way where companies can reach to their potential customer, know them and their behavior (Magnini et al., 2003). Song and Li (2008) analyzed tourism and hospitality literature

published between 2000 and 2007 for modeling tourism demand and identified several data mining techniques that have started to be adopted alongside with traditional models such as the integrated autoregressive moving-average models (ARIMA). From the articles they analyzed, there is a general impression that advanced techniques such as support vector machines outperform traditional ARIMA models, although there is not a single technique that achieves always better results than the others, thus the accuracy is dependent on the specific context and data that defines the problem. However, as Moro and Rita (2016) discussed after analyzing fifty recent articles published between 2013 and 2016, most of the data analysis procedures conducted on tourism and hospitality data are still based on ARIMA models.

As stated previously, a large number of the published research based on customer feedback and, in particularly, in tourism and hospitality, focus on the analysis of the textual contents from users' reviews through techniques based on text mining and sentiment analysis. As an example, Ye et al. (2009b) applied sentiment classification techniques in various online reviews from diverse travel blogs, comparing them with three different supervised machine learning algorithms. In a different line of research, Cao et al. (2011) investigated the impact of online review features hidden in the textual content of the reviews on the number of helpfulness votes of such review texts by applying text mining for extracting the review's characteristics. Still another trend of research includes the application of text mining to the contents of social network sites such as Facebook and Twitter for extracting actionable knowledge, supporting companies on understanding how to perform on a social media competitive analysis, as nowadays it is the number one source for human interaction (He et al., 2013). However, several issues and challenges are brought up when it comes to use text mining. The most widely discussed are context specificities associated with the user and problem being delve, language barriers, and human communication issues such as sarcasm and irony (Aggarwal & Zhai, 2012; Ampofo et al., 2015). For example, many of the reviews published in TripAdvisor are made in each user's native languages. Also, syntactic errors are common on this platform, as users are not concerned with typing errors. Despite some advances in these domains, the intrinsic linguistic subjectivity is still a challenge yet to be overcome. Such difficulty does not exist when only quantitative data based on numerical or categorical features are used for feeding a model based on a data mining technique.

In TripAdvisor, users are able to provide a quantitative rating that generates a score, ranking the property on its overall (O'Connor, 2010). Therefore, the contribution and innovation to the hospitality industry and literature brought by this paper is the application of data mining to all the quantitative features that can be collected from TripAdvisor, in order to model the score given by the reviewers, based on their experience as TripAdvisor users and the hotel's characteristics, instead of the common text mining applied to the written comments published by users.

3. Materials and methods

3.1. Data collection and preparation

After defining the problem, data collection and preparation is the next key step for compiling a dataset that serves as input for modeling. Such dataset is the building block essential for unveiling knowledge through a data mining modeling technique. Moreover, the dataset needs to be composed of a table where each row represents an instance of the problem being addressed and each column represents a feature that characterizes that instance (Witten & Frank, 2005).

Since TripAdvisor owns several domains to cover suffixes from several countries, the data was collected from the TripAdvisor.com website, as the .com is considered the base site where there are reviews belonging to users from every part of the world. Then, it was necessary to filter the information by location, i.e. Las Vegas, Nevada, and more specifically filtering by hotels in the Strip avenue. As a result, a list of 21 different hotels was displayed, allowing to choose a hotel at a time in order to extract the data from each one of them. When opening one of the chosen hotels' pages, access is gained to various information regarding the hotel, such as its address, general quality rating, individual reviews, photos and videos from both the hotel and the previous customers and also the hotel's features. Once the hotel is selected, the procedure undertaken consisted in collecting the data by extracting two reviews per month from the year of 2015, repeating this process for all the 21 hotels. The uniform distribution of the reviews spanned through the different months provides data for building a model that also considers the

seasonality effect known of tourism (Song & Li, 2008). Starting by filtering the time of the year for the period of stay (Dec-Feb; Mar-May; Jun-Aug; Sep-Nov), the search focused on selecting the most completed reviews in order to provide all the information and variables needed until the 24 reviews per year were accomplished. After choosing the reviews, all the features identified from each review, including user characteristics, were collected into a single table, including the score, as it is showed in Figure 1 where each square represents a fragment of data collected. The textual review was also collected, in case it would be needed in future research. The numbers identify the feature extracted enumerated under parenthesis in the column "origin" of Table 1 exhibits the features collected, identified by the "origin" equals to "extracted", with the parenthesized numbering in the same column corresponding to the locations from where each feature was collected, as identified in Figures 1 to 4. The source type groups features into three categories, review features, user features, and hotel features, whereas the data type relates to the types of values that can be assumed by each feature, with categorical type corresponding to a fixed number of enumerated values (e.g., the "gym" feature can assume "yes" or "no") and numerical type corresponding to an ordinal numbered feature. Dates are a particular type of numerical features due to its format restrictions, while "text" type corresponds to unstructured data (here reserved for the "review text").

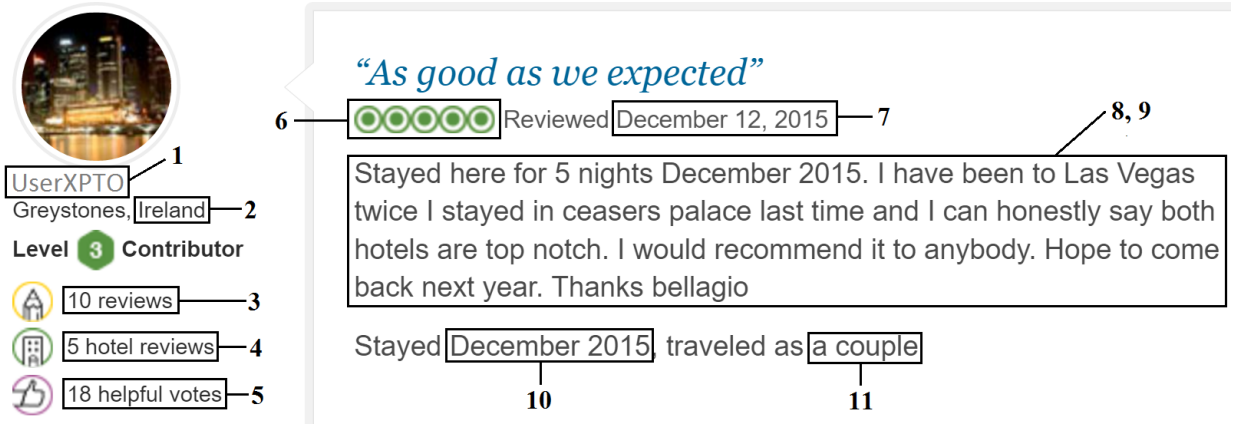

Figure 1 - Review and user features extracted.

For obtaining the date the user has registered in TripAdvisor, it was just needed to pass with the cursor over the username to get such additional information, displayed in Figure 2.

Finally, the webpage with the information supplied by TripAdvisor for each of the 21 hotels was accessed for gathering all relevant features from each hotel (e.g., the link for the Bellagio is: https://www.tripadvisor.com/Hotel_Review-g45963-d91703-Reviews-Bellagio_Las_Vegas-Las_Vegas_Nevada.html). Figure 3 shows a snap-shot of the section where the features from hotel's amenities were extracted, whereas Figure 4 shows the section from where additional relevant features such as hotel's stars and number of rooms were collected.

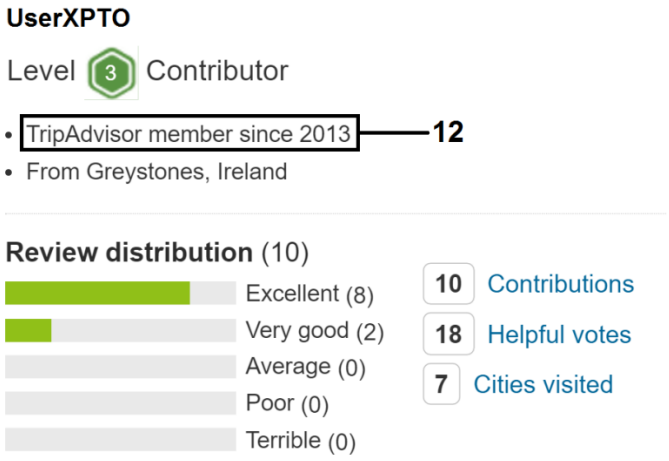

Figure 2 - Extraction of member registered date.

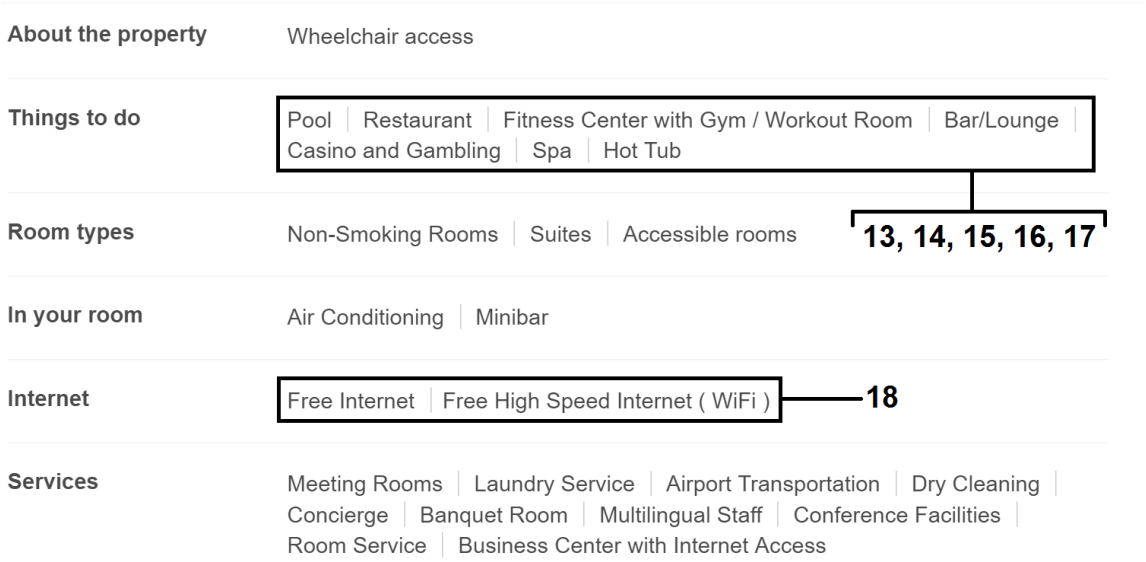

Figure 3 - Extraction of hotel's amenities features.

Figure 4 - Extraction of additional hotel's features.

Table 1 exhibits the features collected, identified by the “origin” equals to “extracted”, with the parenthesized numbering in the same column corresponding to the locations from where each feature was collected, as identified in Figures 1 to 4. The source type groups features into three categories, review features, user features, and hotel features, whereas the data type relates to the types of values that can be assumed by each feature, with categorical type corresponding to a

fixed number of enumerated values (e.g., the "gym" feature can assume "yes" or "no") and numerical type corresponding to an ordinal numbered feature. Dates are a particular type of numerical features due to its format restrictions, while "text" type corresponds to unstructured data (here reserved for the "review text").

Table 1 - List of features.

Feature name	Origin	Source type	Data type	Description	Status
Username	Extracted (1)	User	Categorical	Username as registered in TripAdvisor	Excluded
User country	Extracted (2)	User	Categorical	User's nationality	Included
Nr. Reviews	Extracted (3)	User	Numerical	Number of reviews	Included
Nr. Hotel reviews	Extracted (4)	User	Numerical	Total hotel reviews	Included
Helpful votes	Extracted (5)	User	Numerical	Helpful votes regarding reviews's info	Included
Score	Extracted (6)	Review	Numerical	Review score {1,2,3,4,5}	Included
Review date	Extracted (7)	Review	Date	Date when the review was written	Transformed
Review text	Extracted (8)	Review	Text	Textual content of the review	Excluded
Review language	Extracted (9)	Review	Categorical	Language of the review	Excluded
Period of stay	Extracted (10)	Review	Categorical	Period of stay: {Dec-Feb, Mar-May, Jun-Aug, Sep-Nov}	Included
Traveler type	Extracted (11)	Review	Categorical	{Business, Couples, Families, Friends, Solo}	Included
Member registered year	Extracted (12)	User	Date (year)	Year the user has registered in TripAdvisor	Transformed
Pool	Extracted (13)	Hotel	Categorical	If the hotel has outside pool	Included
Gym	Extracted (14)	Hotel	Categorical	If the hotel has gym	Included
Tennis court	Extracted (15)	Hotel	Categorical	If the hotel has tennis court	Included
Spa	Extracted (16)	Hotel	Categorical	If the hotel has spa	Included
Casino	Extracted (17)	Hotel	Categorical	If the hotel has a casino inside	Included
Free internet	Extracted (18)	Hotel	Categorical	If the hotel provides free internet	Included
Hotel name	Extracted (19)	Hotel	Categorical	Hotel's name	Included
Hotel stars	Extracted (20)	Hotel	Categorical	Hotel's number of stars	Included
Nr. Rooms	Extracted (21)	Hotel	Numerical	Hotel's number of rooms	Included
User continent	Computed	User	Categorical	Continent where the user's country is located	Included
Member years	Computed	User	Numerical	Number of years the user is member of TripAdvisor	Included

Review month	Computed	Review	Categorical	Month when the review was written (from review date)	Included
Review weekday	Computed	Review	Categorical	Day of the week the review was written (from review date)	Included

After the data collection process, the dataset contained 504 records and 21 features extracted (as of "origin=extracted", from Table 1), 24 per hotel, regarding the year of 2015. However, such dataset still needed to be prepared for serving as an input to the modeling stage. Since this data was hand-collected and all the reviews chosen were complete, there were no missing values to be dealt with. However, a closer look at the data allowed to identify a small set of features with few to none value in terms of characterization of each of the reviews in the dataset compiled. These features were excluded from the dataset and are marked accordingly in the column "status" in Table 1. Such is the case for the review language, always in English for the reviews collected; thus, the value remained the same for all the records, meaning it does not provide additional information for characterizing the scores. In fact, most of the reviews found for the Strip's hotels are written in English (e.g., from the 8,878 reviews published on TripAdvisor since ever up to the 31st of July 2016 for the "Encore at Wynn Las Vegas", 7,951 of them are in English, almost 90% of the total), an unsurprising result, given that Las Vegas is in the United States, a native English country with a strong market of domestic tourism (Dawson, 2011) and also the worldwide dissemination of the English language. For the case of the reviews collected, 217 of them are from the United States, 72 from the UK, 65 for Canada, and 36 for Australia, in a total of 390 reviews from native English countries. The username was also excluded, as most of the reviews were from different users (only six of the reviews were made by users from which a previous review was also selected for the dataset). Finally, the textual content of the reviews was not considered for modeling, since it is unstructured and additional techniques would need to be employed, such as text mining. Furthermore, the focus of this research is on knowledge extraction from quantitative features to overcome the limitations of textual reviews mentioned in Section 2, such as the ambiguity of human language.

Another procedure that usually takes place in data mining is feature engineering, which is considered a key step by Domingos (2012). Therefore, a few of the features were transformed (Table 1, "status=transformed") into new ones, which were computed (Table 1,

"origin=computed"). For example, the year when the user registered as a TripAdvisor member is just an occurrence in time, whereas the number of years of membership represents for how long the user is active in TripAdvisor. Thus, the "member registered year" was transformed in "member years". The same happened for "review date", from where "review month" and "review weekday" were computed. Also, the country from where the reviewer is native was used to obtain the corresponding continent, although in this case the "country" feature was kept, since it may conceal meaningful value through user country's characterization of the review score.

The result of these data collection and preparation procedures is a dataset with a total of 19 input features plus the outcome to predict, the score given by users (Table 1 features with status="included").

3.2. Modeling and knowledge extraction

With the dataset ready for modeling, a procedure took place for assessing the robustness of the model built on the data. Figure 5 shows a visual picture of such procedure. The evaluation of the model was executed through a k-fold cross-validation technique where the whole dataset is divided into k folds or sections grouping consecutive reviews from the dataset (Bengio & Grandvalet, 2004). The k value was set to 10 (a value recommended by Refaeilzadeh et al., 2009), implying that 90% (454 reviews) of the data was used for training the model while the remaining 10% (50 reviews) for testing it, thus assuring independence of the split between train and test data. The train-test execution was run 10 times, by varying the fold of data for testing model accuracy, hence computing the predicted score once per record. Since the support vector machine implements a non-linear complex model, to further assure model evaluation, the 10-fold cross-validation was conducted 20 times, with the final score being computed by the average of the 20 executions. Performance modeling was then assessed by computing both MAE and MAPE metrics for these averaged predicted results for each of the reviews in the dataset.

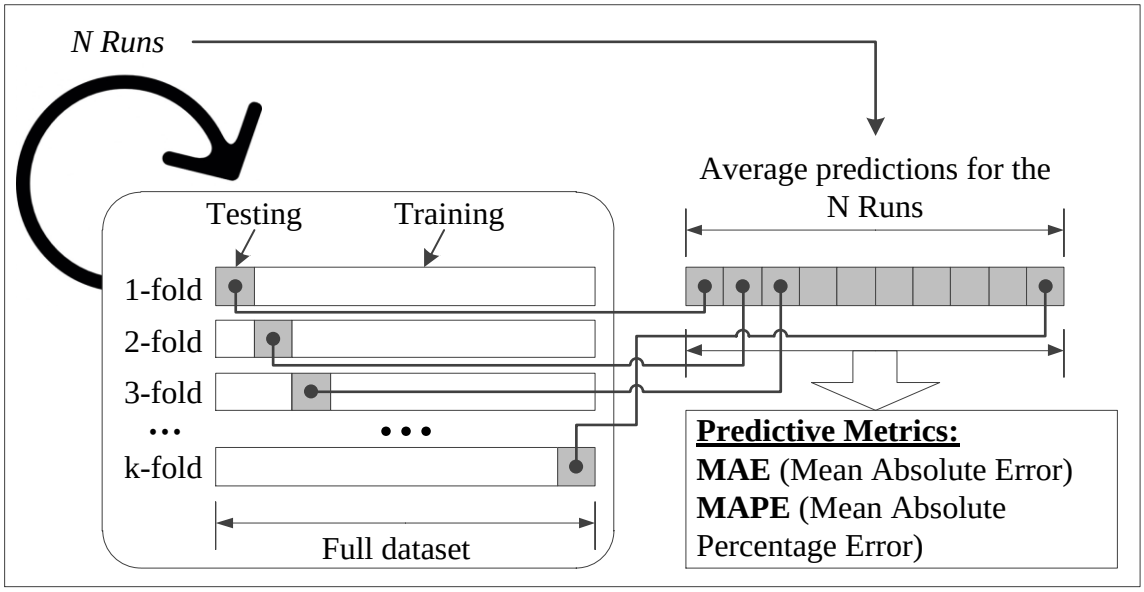

Figure 5 - Modeling performance assessment.

Assuming the input dataset prepared conceals relations between the input features and the score, and that the chosen modeling technique (i.e., support vector machine) is able to unveil such relations, the resulting predictive metrics computed would then comprehend satisfactory results in terms of accuracy. Hence, a model built on the whole dataset using the same modeling technique will also conceal such knowledge, enabling to extract it through the DSA. Figure 6 shows the procedure undertaken for such knowledge extraction. First, to assure similar results, the same metrics (e.g., MAE and MAPE) are computed over the model built on the whole dataset. Then, the same model is used for exposing through DSA which are the features that influence most the score, translating such knowledge in terms of percentage relevance to which each feature contributes for modeling the score. Finally, using also DSA it is possible to observe how each of the most relevant features manages to influence the score.

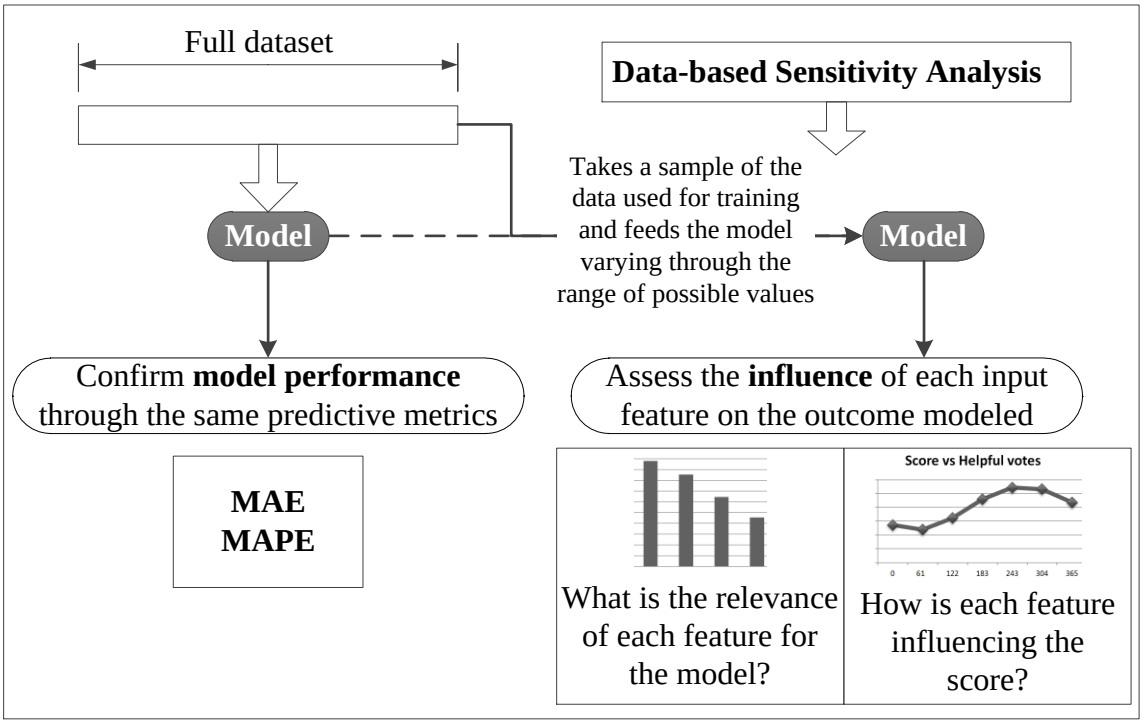

Figure 6 - Knowledge extraction through sensitivity analysis.

To conduct all experiments, the R statistical tool was adopted (see: https://cran.r-project.org/). It provides a free and open source framework with multiple methods and functions to perform data analysis (James et al., 2013). Moreover, it has generated a worldwide enthusiasm translated in a vast community of contributors of a myriad of packages that can be freely downloaded and used for diverse purposes (Cortez, 2014). Specifically designed for data mining, by providing a simple and coherent set of functions, the "rminer" package was chosen (Cortez, 2010). Furthermore, this package also implements functions for extracting knowledge from models through sensitivity analysis, including the DSA.

4. Results and discussion

As described in Section 3 and illustrated in Figure 5, modeling performance was first assessed using an evaluation scheme including a realistic10-fold cross-validation procedure for testing the model with unforeseen data, which was ran twenty times. Table 2 shows the predictions for three

randomly selected reviews with the data used as an input to the model (data is displayed vertically for space optimization purpose only). The predicted score is an average of the 20 executions of the procedure, as described earlier in Section 3. The absolute deviation is the difference between the real and the predicted scores, with the MAE metric resulting from the average of all deviations for the 504 reviews. The percentage deviation corresponds to the relation between the absolute deviation and real score, with the MAPE metric being the computed average of all percentage deviations.

Table 2 - Prediction results for three reviews.

Reviews	**#1**	**#2**	**#3**
User country	USA	USA	Ireland
User continent	America	America	Europe
Member years	2	1	3
Review month	February	October	April
Review weekday	Saturday	Friday	Friday
Nr. Reviews	36	23	19
Nr. Hotel reviews	9	17	9
Helpful votes	25	11	28
Traveler type	Families	Families	Couples
Period of stay	Mar-May	Sep-Nov	Mar-May
Hotel name	Circus Circus Hotel & Casino Las Vegas	Monte Carlo Resort&Casino	Tropicana Las Vegas - A Double Tree by Hilton Hotel
Hotel stars	3	4	4
Nr. Rooms	3,773	3,003	1,467
Free internet	YES	NO	YES
Pool	NO	YES	YES
Gym	YES	YES	YES
Tennis court	NO	NO	YES

Spa	NO	YES	YES
Casino	YES	YES	YES
Real score	**5**	**3**	**5**
Predicted score	**3.9**	**3.6**	**4.6**
Absolute deviation	**1.1**	**0.6**	**0.4**
% deviation	**22.0%**	**20.0%**	**8.0%**

The results for both metrics adopted, MAE and MAPE, can be seen on Table 3. In the scale from 1 to 5 used for the score on TripAdvisor, the support vector machine achieved an average absolute deviation of 0.745, an indicator that it presents a predicted value close to the real score, by less than one. MAPE translates such deviation into a percentage: the average predicted score deviates by 27.32% from the real score. While such results show the model is not totally accurate for every review (as can be seen from the three cases illustrated in Table 2), these also provide proof that the model constitutes a valid approximation for modeling TripAdvisor score. Furthermore, other studies have discovered valid insightful knowledge from a model with a MAPE of around 27% (e.g., Moro et al., 2016b).

Table 3 - Modeling performance assessment metrics.

Metric	**Result**
MAE	0.745
MAPE	27.32%

The knowledge discovery phase aims to provide the major contribution of this research, as it lends insights on the characterization of review scores of such a renowned location as it is the case of Las Vegas Strip, while keeping in mind the relevance widely discussed in the literature of online customers' feedback to the hospitality industry (e.g., Ye et al., 2009a). Thus,

understanding what drives users to publish a given score can ultimately leverage managerial decision support in hospitality. Therefore, the comprehension of the factors that influence why a given hotel is being rated with a certain score can be valuable for managers to act on parameters they control (e.g., hotel related features) and to preventively manage their units according to the expected tourists' demands (e.g., by understanding the more demanding tourists).

As stated previously, the method chosen for knowledge extraction was the DSA. It provides means of presenting for each feature the percentage of relevance that the feature has on the model by analyzing outcome fluctuation to input features' variation. Sensitivity analysis requires a single model, which was built using the whole dataset, as shown in Figure 6. Over this model, the same performance metrics from modeling evaluation were obtained, namely MAE and MAPE, with the results displayed in Table 4. As expected, the values are slightly better than in previous stage, since these represent the predicted values for all data that was also used for training the model, while the 10-fold cross-validation procedure presents a realistic scenario where data used for testing the model was not used for training it, as explained by Moro et al. (2014), who adopted a similar procedure. A MAE result of around 0.5 represents an approximation of the score modeled using the input features, thus providing the needed validation to proceed with knowledge extraction from this model.

Table 4 - Final model performance metrics.

Metric	**Result**
MAE	0.523
MAPE	21.01%

Table 5 exhibits the percentage relevance computed through DSA for all the features while Figure 7 complements it with a visual bar plot of the eleven most relevant features, concealing around 81% of relevance for the model (the remaining eight are indistinctly represented in a single bar labeled "others").

Table 5 - List of features and their relevance.

Feature	**Relevance**
Nr. Hotel reviews	15.0%
Member years	14.1%
Period of stay	10.3%
Nr. Reviews	9.0%
Nr. Rooms	6.1%
Hotel stars	5.1%
Review weekday	5.0%
Helpful votes	4.6%
Traveler type	4.6%
Hotel name	3.8%
User country	3.7%
User continent	3.3%
Free internet	2.7%
Pool	2.6%
Review month	2.5%
Gym	2.5%
Spa	2.4%
Casino	1.8%
Tennis court	0.9%
Total	**100.0%**

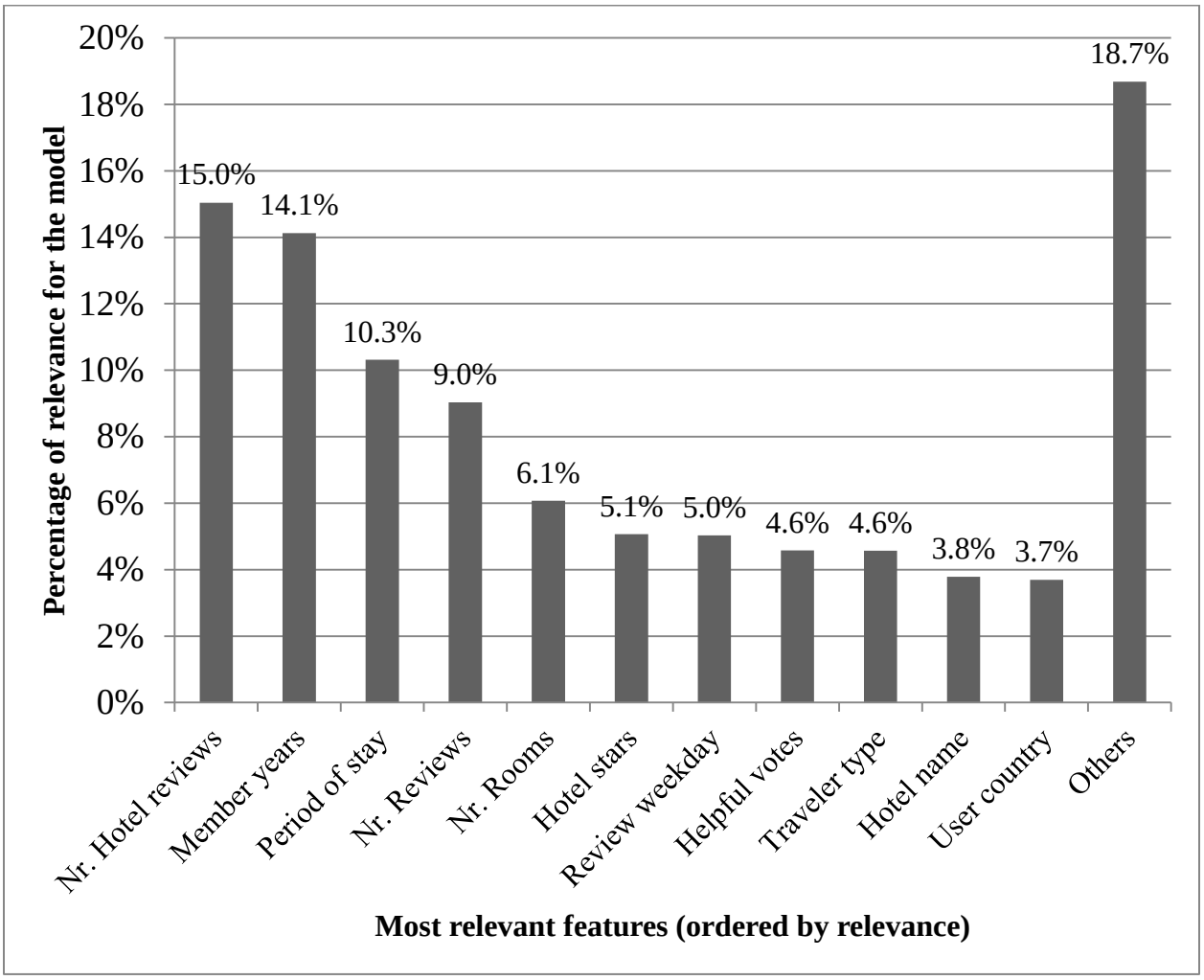

Figure 7 - Most relevant features according to their relevance.

The two most relevant features are both related to the user. The number of reviews of hotels that the user has made contributes with an influence to the final score greater than any of the remaining features, with 15% of relevance. A similar result occurs for the member of years that the user has since first registered in TripAdvisor, with a relevance of 14.1%. In fact, the fourth most relevant feature is the number of reviews, which is closely related to the most relevant feature ("nr. hotel reviews"), as it includes with all the reviews, together with the restaurant and attraction units summing up to hotels' reviews. These three features hold almost 40% of model relevance when modeling the score. This is an interesting discovery, suggesting the score is clearly biased by the users' experience acquired along the time, influencing self-awareness of

what is a fair rate. Hence, managers should have this into account when considering the score their units are having on TripAdvisor.

The period of stay is the third most relevant feature, with 10.3% of influence when compared to the remaining features. Such result was expected, given the seasonality effect known of tourism and hospitality (Song & Li, 2008). Surprisingly, the most relevant hotel features only appear in fifth and sixth places, the number of rooms and stars, respectively. Moreover, previous studies concluded that the number of stars affects online booking (e.g., Ye et al., 2011). Also worth of note is the fact that the weekday the user has published the review plays 5% of the role when it comes to modeling TripAdvisor score. The remaining features are all below 5% in terms of relevance, including hotel name and user country. It was expected that the brand name and image behind the hotel contributed more to user rating, as it is suggested by previous research on hotel brand influence (e.g., Sparks & Browning, 2011). Also worth of noticing is the fact that the features that can be entirely controlled by the hotel, such as the amenities (e.g., free internet, pool, gym, spa, casino and tennis court) are influencing less than 3% each.

Considering the location-based nature of this empirical research, the results hereby presented must be discussed in the light of Las Vegas importance in hospitality and tourism. Las Vegas is a top tourism destination in the United States, which reflects into the high number of reviews in TripAdvisor. As an example, O'Mahony and Smyth (2010) found 146,409 published reviews by 32,002 users prior to April 2009 for Las Vegas, whereas the same study found around half of reviews for Chicago in the same period, a much larger city. These figures reveal that Las Vegas is a very mature tourism market, with its tourists being fully aware of online reviews, whether by publishing new reviews or for obtaining feedback. The more recent study by Rosman and Stuhura (2013) emphasizes the immediacy of online feedback in Las Vegas. In addition, it is known the effect of self-congruity on tourism destinations and, particularly, on Las Vegas tourists (Usakli & Baloglu, 2011). Therefore, experienced tourists translated in a higher degree of TripAdvisor membership may unconsciously be influenced by such experience when providing feedback in such a mature market as Las Vegas. Furthermore, the Las Vegas brand itself is able to generate controversial feelings capable of affecting tourists' perception (Griskevicius et al., 2009). All these characteristics are aligned with the model built on TripAdvisor's review features,

with experience counting as the top influencing factor, while hotel brand having a significant lower relevance.

After analyzing the relevance of features on TripAdvisor score, it is interesting to dive deeper into each of the most relevant ones identified in both Table 5 and Figure 7 in an attempt to understand how these features affect the score rate. Both the most relevant ("nr. Hotel reviews") and the fourth most relevant ("nr. Reviews") features overlap in the sense that the latter includes the former, plus the reviews the user has made on attraction units and restaurants. Therefore, these two features are analyzed together. Figure 8 shows how each influence the score. As expected (Magnini et al., 2003), the experience momentum after the initial first reviews tend to turn the customer more demanding when publishing online score. Nevertheless, such effect is more profound for the global counter of reviews, including attraction units and restaurants. This finding is aligned with previous study by McCartney (2008), which stated that gaming and casino attractions leverage tourists' requirements in terms of hospitality. Hence, global reviews may have the effect of plunging scores to values below 3.9.

Figure 9 displays the effect of the number of years as a TripAdvisor member on the given score. Up to four years of membership, the conclusions are similar to the number of reviews made; however, users registered five years ago or more tend to be more positive by granting better review scores. While for the number of reviews, it can also be observed on Figure 8 a slight increase on the score after a certain threshold (this is particularly visible on the "nr. Reviews" feature), the results for "member years" clearly amplify such tendency, with older members giving scores above new members. Some hypotheses can rise based on this result. One of the most plausible is that tourists with more experience have better knowledge on the destination and units available, thus they will choose the hotels that please them most, resulting in higher scores. Also, experienced TripAdvisor members are probably keener to read other members' reviews and so be better informed for making judged decisions on their own stays (Liu et al., 2015). Nevertheless, more data would be needed to confirm or reject such hypotheses.

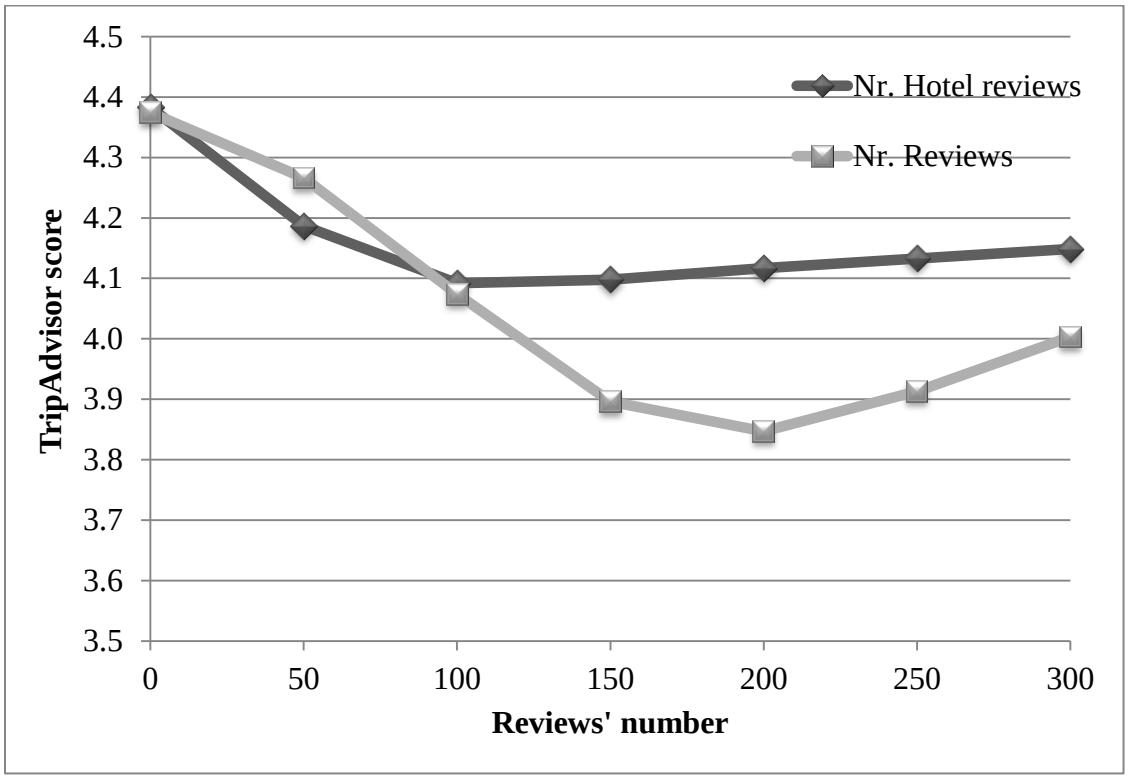

Figure 8 - Influence of "Nr. Hotel reviews" and "Nr. Reviews" on TripAdvisor score.

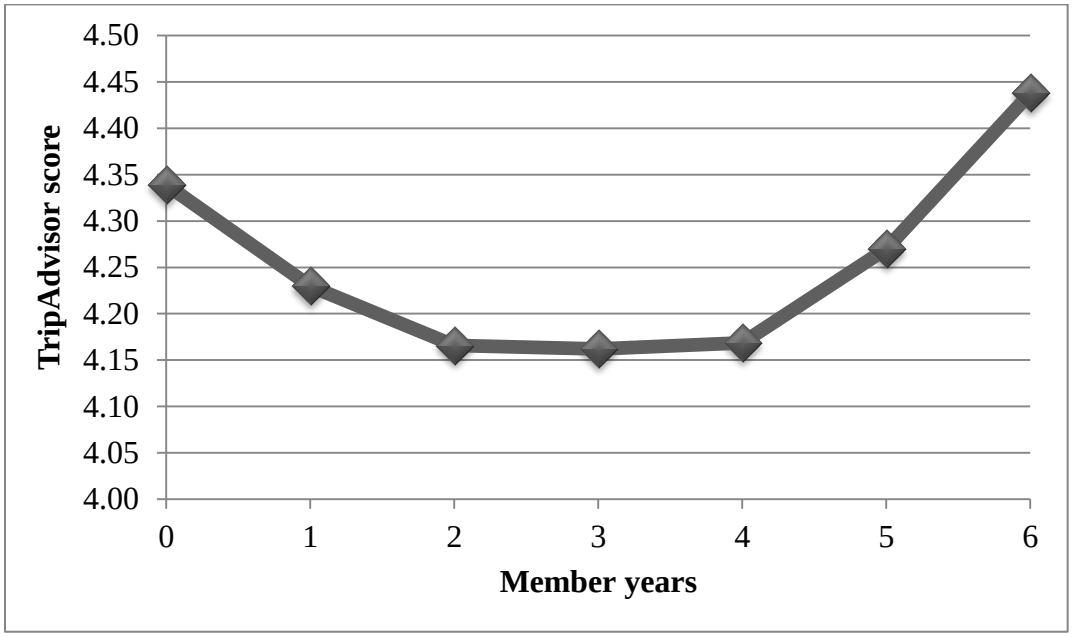

Figure 9 - Influence of "Member years" on TripAdvisor score.

The third most relevant feature for modeling score was the period of stay, in quarter fractions of a year. Figure 10 shows the seasonality effect on TripAdvisor score. Several previous studies are found concluding that Las Vegas holds a seasonality effect on its tourism (e.g., Yang & Gu, 2012; Day et al., 2013). The visible effect on the bar plot is very small, almost negligible, with Sep-Nov reaching the peak of 4.37 of score, while Mar-May bottoms at 4.30. Nevertheless, by holding relevance above 10% for the model implicates its variation although small does affect TripAdvisor score and probably such influence gets amplified in aggregation with the remaining features.

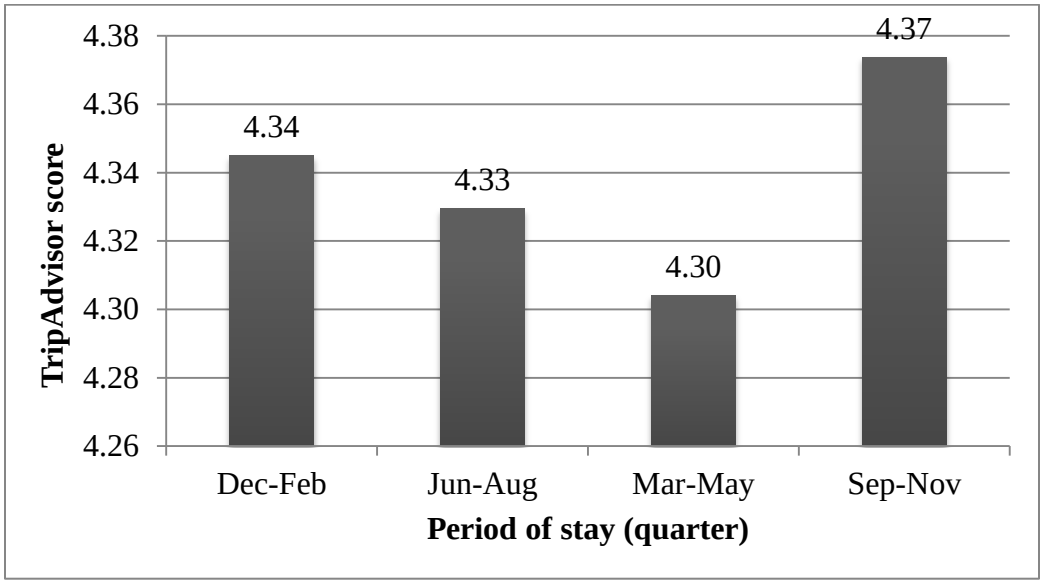

Figure 10 - Influence of "Period of stay" on TripAdvisor score.

The number of rooms the hotel unit has is the fifth most relevant feature, although with a contribution of just 6.1% pales in comparison with the top four, all above 9% of relevance. Still, it is the most relevant feature in respect to hotel specifications. Figure 11 shows that smaller units tend to have better review scores. This effect is significant, with the average difference score between an hotel with 200 rooms and another with 3,800 reaching 0.4 points. The recent study by Jiménez et al. (2016) based on Spain and Portugal hotel units also found a similar relation: as the number of rooms increases, the TripAdvisor score decreases. Hotels smaller tend to offer a friendlier and non-crowd environment which may be promoted as an advantage against large resorts, suiting better tourists enjoying quiet stays inside the unit (Chambers, 2010).

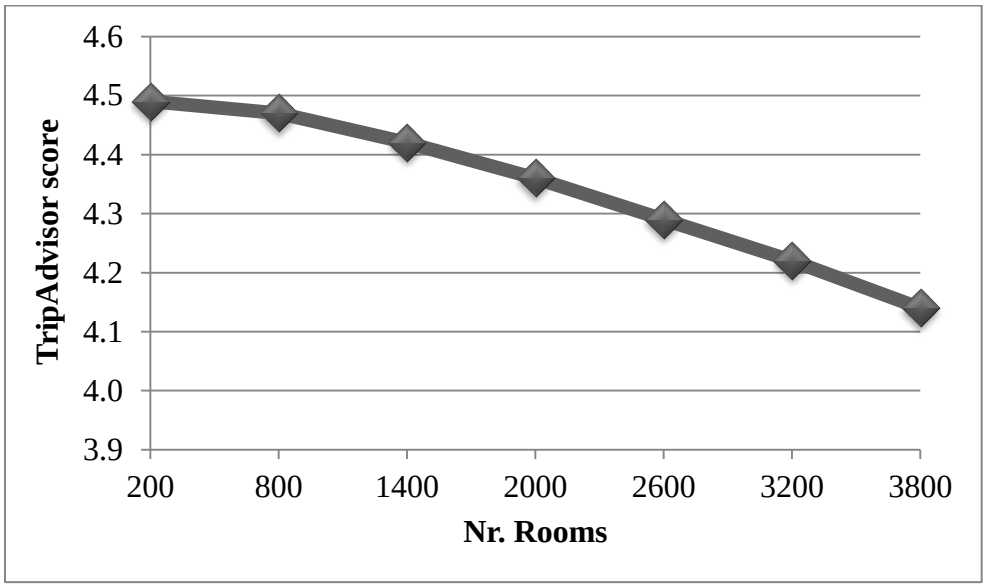

Figure 11 - Influence of "Nr. Rooms" on TripAdvisor score.

Figure 12 displays the effect of the number of stars of the hotel on TripAdvisor score. The result is expected: the higher the number of stars, the higher the score. Las Vegas Strip hotels' range from three to five stars. Hu and Chen's (2016) study is aligned with the findings unveiled from Figure 12 in that hotel stars influence positively reviews' ratings.

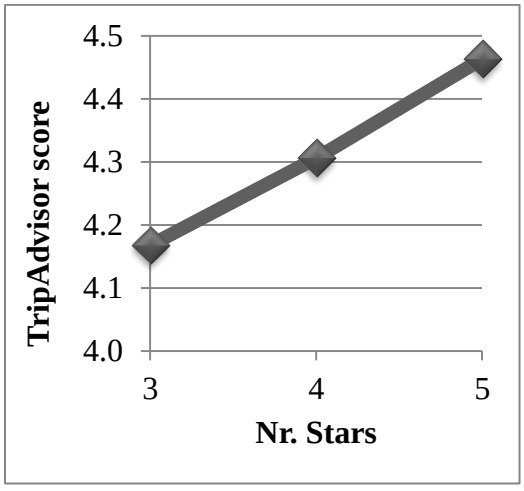

Figure 12 - Influence of "Nr. Stars" on TripAdvisor score.

The seventh most relevant feature is a surprise: the weekday when the review was published achieved a relevance of 5% (Figure 7). From Figure 13 it is possible to observe that the weekday

influences directly TripAdvisor score in a range of 0.24 points (from 4.24 on Tuesday to 4.48 on Saturday). The effect of seasonality is known in tourism, but the finding related to the influence of the weekday's of publication has no precedent in tourism. Furthermore, user feedback may vary a lot in terms of lag related to the period of stay, as some tourists provide feedback directly on sight, while others wait some days before writing the review. Nevertheless, other studies on social media have also found an influence of the weekday of publication on the impact of publishing contents, such as the finding by Moro et al. (2016b) on a company's Facebook posts. Seemingly reviews published near the weekend tend to receive better scores, as shown in Figure 13.

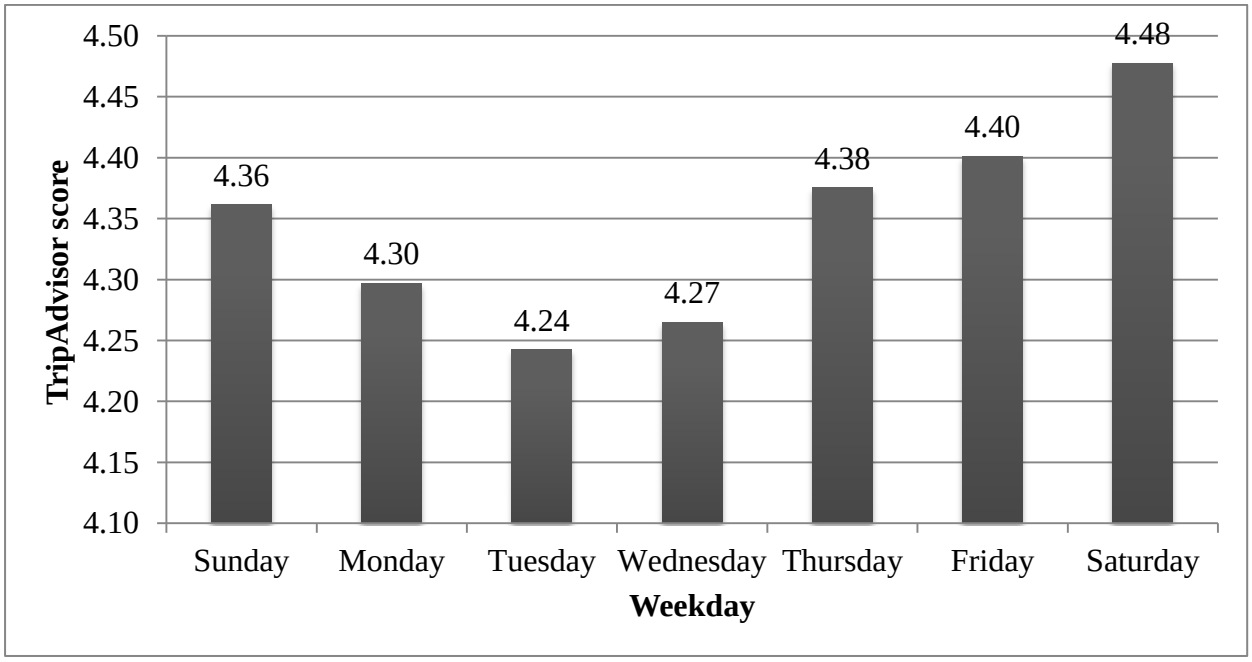

Figure 13 - Influence of "Weekday" on TripAdvisor score.

Other features contributing with a relevance below 5% including "helpful votes", "traveler type", "hotel name" and "user country" are not scrutinized in this paper. Nevertheless, each of them plays a role on the model built, although its lesser relevant role in comparison with the top influencing features.

5. Conclusions

It is currently unquestionable that online feedback reviews in tourism have the power to influence to a certain degree forthcoming tourists. Hence, hospitality unit managers have recently included such source of information in their decision making processes. TripAdvisor is the largest online platform for providing feedback on tourism and hospitality and one of the main sources for managers to control customer feedback.

A TripAdvisor member has mainly two means for providing feedback: a free text area for input of textual comments; and a quantitative score between 1 and 5. The textual comments, by concealing interesting user sentiments, have been widely studied in the literature. However, knowledge extraction based on such comments is usually harder to achieve when compared to the quantitative score. Furthermore, the inherent subjectivity associated with human language poses difficult challenges to overcome. On the opposite side, the quantitative score is an objective measure, easier to model. Still, research on the score is rather scarce in comparison to research on textual reviews. Hence, the knowledge extraction procedure presented in this paper is based on modeling TripAdvisor score. What are the characteristics underlined in a review that ultimately lead to a certain score? And how are these features influencing a tourist to rank a stay in a hotel unit? The present study aimed at enlightening hospitality research by answering these questions for the case of Las Vegas Strip.

For the empirical research presented in this paper, data from reviews was collected for the Strip avenue in Las Vegas, which is a very mature location-based market linked to gaming and pleasure industries, translated in a high number of reviews on TripAdvisor for each of its 21 hotel units. A total of 21 quantitative features such as the period of stay, the number of reviews the user had previously made on TripAdvisor, and the hotel number of stars, were chosen for extraction. For each of the 21 hotel units, two reviews per month published in 2015 were then extracted, resulting in a dataset with 504 reviews. Then, a data preparation procedure took place for setting up data to serve as an input for modeling the score. The result is a final dataset with 19 tuned features for mining knowledge.

With the dataset compiled, data mining could then occur using such set of reviews' data for modeling TripAdvisor score. The technique chosen for modeling was the support vector machine, which is a recent method developed in the 1990's that has been successively adopted for numerous regression problems. The experimental approach for building the model comprised two main stages: first, for assessing modeling performance, a realistic 10-fold cross-validation procedure was run for 20 times, with the prediction results computed as the average for the predicted scores in each of the 20 executions. A MAE of 0.745 and a MAPE of 27% assured the deviation from the score predicted and the real value constituted an interesting approximation as a predictive model. Nevertheless, the main goal was to extract useful knowledge by understanding how each of the features influenced the review outcome. Thus, the same support vector machine experimental setup was applied to the whole dataset for building a model that effectively mimicked users' characteristics and ratings. Such procedure constituted the second stage of the approach. A data-based sensitivity analysis was applied over the model to understand to which measure the features affected the score. The result is a rank of features by its percentage of relevance to the model measured by the sensibility of the outcome score in terms of its deviation to the variation of the input features.

The knowledge unveiled shows the experience of the user as a TripAdvisor member plays a key role when publishing a score on TripAdvisor. The number of hotel reviews ranks top of the relevance list, with 15% of relevance, while the number of reviews (which includes hotels, attractions and restaurants) ranks fourth on the list, with 9%. Seemingly, the more reviews the user publishes, the less is the score granted, meaning experienced users tend to be more demanding. The number of years of TripAdvisor membership is the second most relevant feature, with 14.1% of relevance. Users with 2 to 4 years as members tend to rate worse scores, while older users happen to grant better scores after 4 years. Most likely more informed users tend to make better judged choices; hence scores also tend to improve. The third most relevant feature is the period of stay, revealing an expected result, given the seasonal nature known of tourism. Still, the differences between stays within each quarter are small. The number of rooms of the hotel and the stars, with relevances of 6.1% and 5.1% respectively, are the most relevant hotel characteristics. This is an interesting finding, as it shows the influence of hotel specifications is rather smaller, when compared to user self-awareness and keenness on his/her own stay

experience. A hotel with more stars tends to have better scores, while the effect is the opposite when it comes to hotel size in terms of number of rooms, i.e., a smaller hotel gets better scores.

It should be noted that, by being a location-based study, users' awareness of Las Vegas brand itself must be an accountable factor on influencing score. Furthermore, such renowned brand is able to generate controversial feelings capable of affecting tourists' perception. This fact may also play a role on the lower ranked hotel features in terms of relevance when compared to user characteristics. As Magnini et al. (2003) discussed, customer satisfaction may bias a data mining approach in tourism due to the relative importance each user attributes to certain characteristics. The present study sheds additional light by concluding that experience as a TripAdvisor member does affect the score rank given by users. However, the present study is focused solely on reviews for hotels in Las Vegas Strip, thus its conclusions have to remain location-based. Therefore, additional research is in demand to confirm or refute the possible generalization of TripAdvisor experience influence on score. Moreover, future research may include studying different locations, with different characteristics. Also, more features from other sources may be included in the model, if available.

References

AleEbrahim, N., & Fathian, M. (2013). Summarising customer online reviews using a new text mining approach. International Journal of Business Information Systems, 13(3), 343-358.

Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.

Ampofo, L., Collister, S., O'Loughlin, B., & Chadwick, A. (2015). Text Mining and Social Media: When Quantitative Meets Qualitative and Software Meets People. Innovations in Digital Research Methods, 161-91.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research, 5, 1089-1105.

Breazeale, M. (2009). An Assessment of Electronic Word-of-Mouth Research. International Journal of Market Research, 51(3), 297-318.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems, 50(2), 511-521.

Chambers, L. (2010). Destination competitiveness: An Analysis of the characteristics to differentiate all-inclusive hotels & island destinations in the Caribbean. Thesis. Rochester Institute of Technology.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: a study of consumer perceptions. Electronic Commerce Research and Applications, 2(3), 203-215.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. In Industrial Conference on Data Mining (pp. 572-583). Springer Berlin Heidelberg.

Cortez, P. (2014). Modern optimization with R. New York. Springer.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17.

Dawson, M. (2011). 'Travel Strengthens America'? Tourism promotion in the United States during the Second World War. Journal of Tourism History, 3(3), 217-236.

Day, J., Chin, N., Sydnor, S., & Cherkauer, K. (2013). Weather, climate, and tourism performance: A quantitative analysis. Tourism Management Perspectives, 5, 51-56.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. Decision Support Systems, 45(4), 1007-1016.

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. Information and Communication Technologies in Tourism 2008, 35-46.

Griskevicius, V., Goldstein, N. J., Mortensen, C. R., Sundie, J. M., Cialdini, R. B., & Kenrick, D. T. (2009). Fear and loving in Las Vegas: Evolution, emotion, and persuasion. Journal of Marketing Research, 46(3), 384-395.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464-472.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? Journal of Interactive Marketing, 18(1), 38-52.

Hu, Y. H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. International Journal of Information Management, 36(6), 929-944.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York. Springer.

Jiménez, S. M., Morales, A. F., de Sandoval, J. L. X., & Stefaniak, A. C. (2016). Hotel assessment through social media–TripAdvisor as a case study. Tourism & Management Studies, 12(1), 15-24.

Lau, K. N., Lee, K. H., & Ho, Y. (2005). Text mining for the hotel industry. Cornell Hotel and Restaurant Administration Quarterly, 46(3), 344-362.

Lee, K. (2015). Transforming for the Future: The New Economic Driver for the Las Vegas Tourism Industry (Master Thesis, University of Nevada, Las Vegas, United States). Retrieved from http://digitalscholarship.unlv.edu/thesesdissertations/2611/

Liu, Z., Le Calvé, A., Cretton, F., Balet, N. G., Sokhn, M., & Délétroz, N. (2015). Linked Data Based Framework for Tourism Decision Support System: Case Study of Chinese Tourists in Switzerland. Journal of Computer and Communications, 3(05), 118-126.

Magnini, V. P., Honeycutt Jr, E. D., & Hodge, S. K. (2003). Data mining for hotel firms: Use and limitations. Cornell Hospitality Quarterly, 44(2), 94.

Mauri, A. G., & Minazzi, R. (2013). Web reviews influence on expectations and purchasing intentions of hotel potential customers. International Journal of Hospitality Management, 34, 99-107.

Mazurek, G. (2009). Web 2.0 implications on marketing. Management of Organizations: Systematic Research, 51, 69-82.

McCartney, G. (2008). The CAT (casino tourism) and the MICE (meetings, incentives, conventions, exhibitions): Key development considerations for the convention and exhibition industry in Macao. Journal of Convention & Event Tourism, 9(4), 293-308.

Min, H., Min, H., & Emam, A. (2002). A data mining approach to developing the profiles of hotel customers. International Journal of Contemporary Hospitality Management, 14(6), 274-285.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

Moro, S., Cortez, P., & Rita, P. (2016a). A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features. Neural Computing and Applications, In press.

Moro, S., Rita, P., & Vala, B. (2016b). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), 3341-3351.

Moro, S., & Rita, P. (2016). Forecasting tomorrow's tourist. Worldwide Hospitality and Tourism Themes, In press.

Neirotti, P., Raguseo, E., & Paolucci, E. (2016). Are customers' reviews creating value in the hospitality industry? Exploring the moderating effects of market positioning. International Journal of Information Management, In press.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592-2602.

O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. Journal of Hospitality Marketing & Management, 19(7), 754-772.

O'Mahony, M. P., & Smyth, B. (2010). A classification-based review recommender. Knowledge-Based Systems, 23(4), 323-329.

O'Reilly, T., & Battelle, J. (2009). Web squared: Web 2.0 five years on. O'Reilly Media, Inc.

Palmer, A., Montaño, J. J., & Sesé, A. (2006). Designing an artificial neural network for forecasting tourism time series. Tourism Management, 27(5), 781-790.

Papathanassis, A., & Knolle, F. (2011). Exploring the adoption and processing of online holiday reviews: A grounded theory approach. Tourism Management, 32(2), 215-224.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In Encyclopedia of database systems (pp. 532-538). Springer US.

Ro, H., Lee, S., & Mattila, A. S. (2013). An affective image positioning of Las Vegas hotels. Journal of Quality Assurance in Hospitality & Tourism, 14(3), 201-217.

Rosman, R., & Stuhura, K. (2013). The implications of social media on customer relationship management and the hospitality industry. Journal of Management Policy and Practice, 14(3), 18-26.

Rowley, R. J. (2015). Multidimensional community and the Las Vegas experience. GeoJournal, 80(3), 393-410.

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. Journal of Travel & Tourism Marketing, 32(5), 608-621.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. Tourism Management, 29(2), 203-220.

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. Tourism Management, 32(6), 1310-1323.

Tinoco, J., Correia, A. G., & Cortez, P. (2011). Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time. Construction and Building Materials, 25(3), 1257-1262.

Sharda, R., Delen, D. & Turban, E. (2015). Business Intelligence and Analytics: Systems for Decision Support, 10th edition. Pearson Education.

Usakli, A., & Baloglu, S. (2011). Brand personality of tourist destinations: An application of self-congruity theory. Tourism Management, 32(1), 114-127.

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. Tourism Management, 30(1), 123-127.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Yang, L. T., & Gu, Z. (2012). Capacity optimization analysis for the MICE industry in Las Vegas. International Journal of Contemporary Hospitality Management, 24(2), 335-349.

Ye, Q., Law, R., & Gu, B. (2009a). The impact of online user reviews on hotel room sales. International Journal of Hospitality Management, 28(1), 180-182.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human Behavior, 27(2), 634-639.

Ye, Q., Zhang, Z., & Law, R. (2009b). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications, 36(3), 6527-6535.