University Institute of Lisbon

Department of Information Science and Technology

Parking Guiding System with Occupation Prediction

Gonçalo Pereira Alface

A Dissertation presented in partial fulfillment of the Requirements for the Degree of
Master in Computer Engineering

Supervisor
PhD. Professor João Carlos Ferreira, Assistant Professor
ISCTE-IUL
Co-Supervisor
PhD. Professor Rúben Filipe Pereira, Assistant Professor
ISCTE-IUL

September, 2019

Resumo

A disponibilidade de estacionamento é um recurso cada vez mais escasso e caro nas grandes cidades, e este problema é considerado um dos mais críticos nos sistemas de gestão de transportes dentro de uma grande cidade. Para abordar este problema, uma prova de conceito é apresentada como uma forma de guiar um condutor para o parque de estacionamento com lugares disponíveis através de um processo de previsão usando dados passados, correlacionados com o tráfego, condições climáticas e características do período de tempo (ano, mês, dia, feriados, e assim por diante).

Uma seleção de características foi realizada pelo estudo de padrões de dados, a fim de entender a afluência do estacionamento e como certas características os influenciam, bem como para compreender as mudanças repentinas na ocupação total do estacionamento e quais características realmente importam e têm um impacto sobre a ocupação total. Essas conclusões ajudaram a criar um modelo preditivo robusto e eficiente a fim de prever a taxa de disponibilidade do estacionamento com mais precisão.

Três algoritmos foram usados para construir os modelos preditivos como forma de testar o mais eficiente e preciso, a saber: Gradient Boosting Machine, Decision Random Forest e Neural Networks. Foram também testados vários tipos de modelos com o objetivo de melhorar os resultados obtidos, bem como compreender o impacto de cada um dos processamentos de dados utilizados.

Para complementar, foi criado um algoritmo de decisão para orientar o condutor para o parque de estacionamento mais indicado e que apresente melhores condições, tendo em conta a localização e as características do condutor, como o mais provável de ter um lugar de estacionamento disponível, mais próximo da posição atual do utilizador ou um preço mais atrativo para o condutor. Finalmente, estes desenvolvimentos são integrados numa aplicação móvel de forma a que o utilizador consiga aceder através de uma interface.

Palavras-chave: disponibilidade de estacionamento, previsão, aplicação móvel, probabilidade, gestão de estacionamento

Abstract

Parking availability is an increasingly scarce and expensive resource within large cities, and this problem is considered to be one of the most critical transportation management system inside a big city. To approach this problem a proof of concept is presented as a way to guide a driver to the possible free parking lot through a prediction process using past data, correlated with traffic, weather conditions and time period features (year, month, day, holidays, and so on).

A feature selection was performed by the study of data patterns, in order to understand the parking lot affluence and how certain features influence them, as well as to comprehend the sudden changes in the total occupation of the parking lot and which features really matter and have an impact on the total occupation. Those conclusions helped to create a robust and efficient predictive model in order to predict the parking lot availability rate more accurately.

Three algorithms were used to build the predictive models as a way to test the most efficient and accurate one, namely Gradient Boosting Machine, Decision Random Forest and Neural Networks. Various types of models were tested with the aim of improving the results obtained, as well as understanding the impact of each of the processing of the data used.

To complement this, a decision algorithm was created to guide the driver to the most optimal parking lot that presents better conditions, taking into account the location and driver characteristics, like the park more likely to have an available parking space, closer to the user's current position or a more attractive price for the driver. Finally, these developments are integrated into a mobile application in order to work like an interface that the driver can interact.

Keywords: parking availability, prediction, mobile app, probability, parking management

Acknowledgements

I would like to express a big thanks to my supervisors, Professor João Carlos Ferreira and Professor Rúben Pereira, for their guidance, support and feedback, as well as encouraging me to take this opportunity.

A special and huge thanks to my family, for educating and motivating me to achieve better and higher goals and for believing in me wholeheartedly. To my girlfriend, Mariana, for being the most generous, loving and understanding person ever and for believing in me, more than I do myself. And lastly, to my friends for the encouragement, the laughs and for helping me to keep my sanity. Once again, thank you all for the encouragement, for believing and for the unconditional support and love. None of this would be possible without them.

Contents

List of Figures

List of Tables

Abbreviations

PGIS Parking Guidance and Information System

IPR Intelligent Parking Reservation

CRISP-DM Cross-Industry Standard Process for Data Mining

IoT Internet of Things

SVR Support Vector Regression

RMSE Root Mean Square Error

NN Neural Networks

API Application Programming Interface

DR Duration of the Route

DPD Distance from the Parking lot to the final Destination

PPH Price Per Hour

AR Availability Rate

JSON JavaScript Object Notation

CSV Comma-Separated Values

DRF Distributed Random Forest

GBM Gradient Boosting Machine

TP True Positive

TN False Negative

FP False Positive

FN False Negative

MOJO Model ObJect, Optimized

ISCTE-IUL Instituto Superior de Ciências do Trabalho e da Empresa - Instituto Universitário de Lisboa

Chapter 1

Introduction

In this chapter the scope of this work is introduced by giving some context and motivation, as well as the objectives and the methodology used. At the end of the chapter the structure of the document is also introduced.

1.1 Motivation and Overview

Nowadays, if drivers want to leave home and go to a desired destination, they can use the vehicle's navigation system or third party applications to find the destination with ease. Those systems give clear information about the time it takes to arrive to the destination, but upon reaching it the driver needs to find an available parking space which can take quite some time and effort (Klappenecker, Lee, & Welch, 2014). This phenomenon is intensified when people who do not know the city, such as tourists or non-residents, are looking for a parking space, as residents have a better knowledge of the place. We have all been in the situation of trying to locate a free parking space to park the car, and after minutes and minutes of searching, we start to get frustrated and our stress levels increase, making us angrier and therefore increasing the probability of making an error, possibly causing an accident (Ionita, Pomp, Cochez, Meisen, & Decker, 2018). The strategy used by most of the drivers looking for a free parking space is called

"Blind Search" (Shin & Jun, 2014) and is used by the drivers when there is no information given regarding the current status of the parking lot. This strategy is based on the driver going around the park looking for an empty parking space until they find a free one.

One of the major contributors to city traffic is the search for parking spaces (Bock, Martino, & Origlia, 2017), being responsible for 30% of the total traffic flow in dense urban scenarios (Shoup, 2006). Most of the time this increase is due to the drivers not having knowledge about where the parking lots are and if they have available parking spaces matching their expectations, forcing users to roam around. People looking for a free parking space often double park which is, by itself, an illegal activity (Giuffrè, Siniscalchi, & Tesoriere, 2012) or even resort to unauthorized spaces (Gantelet & Lefauconnier, 2006) resulting in a major contribution to congestion on the roads, and other impacts like: increase the danger and probability of accidents involving pedestrians as cars block sight lines, since the drivers looking for a free parking space are more distracted and more stressed being more prone to cause accidents either with pedestrians or other vehicles (Ionita et al., 2018), degradation at the quality of public transportation, the time a driver wastes, increase of noise (Shin & Jun, 2014), safety concerns for motorists and cyclists performing maneuvers around double-parked cars, increase of the pollution levels inside the city for up to 40% (Pflügler, Köhn, Schreieck, Wiesche, & Krcmar, 2016) and unnecessary use of fuel (Klappenecker et al., 2014), increasing carbon emissions up to 27% and increase of 54% of time wasted in traffic (Giuffrè et al., 2012; Tayade & Patil, 2016).

Vehicles where initially invented to increase convenience and comfort in people's every day life, however the demand for a parking space increases progressively over time, where a driver in the first 15 minutes of demand will look for an available spot within a distance of less than 200 metres from his final destination, but by exceeding 15 minutes of demand, this demand range increases, sometimes exceeding 500 metres (Gantelet & Lefauconnier, 2006), this being a major inconvenience for drivers. Parking is becoming an expensive but also scarce resource being regarded as one of the major issues in city transportation management since spatial resource of a city is limited and the construction of new parking spaces is expensive, sometimes due to the public transport policy, in almost any major city in the world (Giuffrè et al., 2012). With the increase of the Smart Cities, which aim to make a city more efficient and improve the lives of its citizens, and that about 70% of the world population will start living in cities and surroundings by 2050 (Tayade & Patil, 2016), this is one of the major problems to tackle in the coming years (Chen, Pinelli, Sinn, Botea, & Calabrese, 2013).

A report on 4 districts in France (Gantelet & Lefauconnier, 2006) concluded that 64% of interviewed residential car owners have abandoned their trips for not finding an available parking space and estimated that 70 million hours are spent each year in France looking for parking spaces, resulting on a total of 700 million euros lost each year. Other study done on the parking situation in Schwabing (Germany) concluded that the annual total economy damage due to traffic caused by the search of an empty parking space had been estimated as much as 20 million euros (Caliskan, Barthels, Scheuermann, & Mauve, 2007). On average, it takes 12 minutes for a driver to find a free parking place (Pflügler et al., 2016) and a nationwide survey done in Netherlands says that if employer-provided and residential parking are excluded, a total of 30% of car trips end with the search for a free parking space (Bock & Sester, 2016). In the United States of America, a car looking for a free parking space in Los Angeles needs to go around a block at least two and a half times to find a clear space to park, adding a total of around 1,500,000 excess kilometers traveled, resulting on almost a total of 178,000 liters of gas wasted and a total of 730 tons of carbon dioxide produced in one year (Klappenecker et al., 2014).

It is important to define parking availability to be the remaining parking spaces in a parking lot, and as of what was said earlier, parking availability is among the most important factors affecting car-based trip decisions and traffic conditions in urban areas. Drivers' decisions are influenced by past experience, as well as real-time (on road) perceptions (E. I. Vlahogianni, Kepaptsoglou, Tsetsos, & Karlaftis, 2016), meaning that parking is such a case where prior knowledge on possible prevailing conditions (e.g. difficulty in finding a parking space, parking costs, and so on) affects drivers' parking decisions, just like the knowledge of current conditions (e.g. day of the week, if it is raining and how much, temperature, events around the parking spaces, and more) affects parking availability (Rong, Xu, Yan, & Ma, 2018). If the city has means to inform the drivers in advance about the availability of parking spaces at and around their intended destination, the traffic congestion can be efficiently controlled (Zheng, Rajasegarar, & Leckie, 2015). The possibility of knowing the park availability in advance affects private car based trip decisions, since if a driver knows that he has a parking place available, he goes there or if there are no spaces available, he can rethink his route and go to another place with a space available, thus avoiding turning around and wasting time and fuel (E. I. Vlahogianni et al., 2016).

Services to mitigate the parking search problem and give parking information to the public has increased recently (Xiao, Lou, & Frisby, 2018). These systems can reduce the queues in front of parking lots and reduce the number of kilometres the driver travels, thus reducing the amount of fuel used and in turn reducing the amount of carbon sent to the atmosphere (E. I. Vlahogianni et al., 2016). One type of service is the Parking Guidance and Information System (PGIS), being an effective way to enhance city parking management, allowing the balance of vehicle parking in the different parking lots, where vehicles can be guided to an emptier parking lot, thus resulting in a better parking management (Z. Wang, Yi, Liu, & Zhang, 2007). Services like this can be very useful to resolve the problem users have while roaming around looking for an available parking space (Bock et al., 2017), but does need detailed real-time parking availability, usually collected by sensors, which sometimes prove quite expensive and hard to maintain, or by crowd-sensing solutions like mobile applications or probe vehicles, which are often difficult to obtain and small in quantity, not being able to have a complete real-time parking availability data (Bock & Sester, 2016), since the usage of roads is highly irregular where the main roads have a regular coverage, but secondary streets are not that often sensed. However, this type of systems do not take into account the time it takes to reach the parking lot, meaning that the current situation on the parking lot will not necessarily be the same upon reaching it (Caicedo, Blazquez,

& Miranda, 2012), which, in turn, can result in the problems that these systems are trying to solve.

Other services involve mobile payment, such as smartphone use, and call-ahead reservation services (Xiao et al., 2018), called Intelligent Parking Reservation (IPR), where the users can reserve a parking space at the final destination before starting the trip. Those systems are able to interact with the navigation systems from the vehicles, as well as Internet users, giving access to real-time information about the parking lot, as well as information on the characteristics of the park, from capacity, hourly price, location, etc. This information can then be shown in the PGIS or even in the IPR system, thus being able to guide drivers to a free parking space during their trips or before their departures, reducing the time spent parking, and in the case of the IPR, paying the fee in advance, avoiding queues (Caicedo et al., 2012). However, this call-ahead reservation could lead to high reservation costs, because the driver can get stuck in traffic or change his destination midway, having to pay even if the car is not inside the infrastructure.

Predictive parking information reveals to be a very useful information tool for all drivers, as users will make informed choices, improving and optimizing parking searching in a way that people could start to plan their route depending on the availability of parking spaces at the destination upon reaching it. If systems based on the prediction of the parking occupancy are well managed and can produce accurate and real occupancy values then it can enhance the moment of decision for the users (Caicedo et al., 2012). According to (Caicedo, Robuste, & Pita, 2006), users who have information about the availability rate of the parks have 45% more success in their decisions than those who do not have access to any information. The way to achieve parking availability foresight is a big challenge, as the system to be developed needs to generate precise parking availability values, because a system that underestimates free parking spaces will not forward the users to that parking lot and if the system returns more free parking spaces than there really are, it would forward a user to a parking lot with no free parking spaces, revealing a problem for the user and smear the confidence on the system (Richter, Martino, & Mattfeld, 2014). Other challenges with those types of systems are the interaction between the parking lots in an area, and how user behaviors affect the parking availability (Zheng et al., 2015).

Smart parking guidance based on the real-time information is still in its infancy, but considering that the parking problem is one of the most complicated issues in a big city and its limited availability can cause problems to the quality of life of people and the deterioration of urban mobility, it is important to improve the current parking guidance solutions and use them in an intelligent and more efficient way (Shin & Jun, 2014). Road congestion, parking availability and more, are problems that with the help of smart cities, will be better dealt with, so a system like the one proposed in this study is more and more needed, where we need to create a parking availability prediction system that will help people plan their trips ahead of time, reducing traffic congestion and, therefore, save time and fuel and keep transport systems more efficient and roads safer.

1.2 Objectives

The present work tackles the increasing problem of parking availability inside big cities, by proposing a system to give accurate real-time advice and guidance in a mobile App about parking availability for drivers. This objective was divided into three sub-objectives to be better dealt with and can be viewed as follows:

- Objective 1: Identify the most important features to predict the parking availability in closed parks;
- Objective 2: Identify the most suitable predictive model;
- Objective 3: Develop a decision algorithm to choose the optimal park considering several factors.

For the first objective there is a need to understand how the parking availability patterns occur throughout the days, so one of the focus in this research is to understand the occupation patterns of a parking lot over time, and which external and internal park features most influence the occupation rate of a park. Understanding all the elements that influence the parking lot occupancy allows us to create a more efficient and robust solution, so that cyclical and unexpected factors can be dealt with the best way possible.

The second objective focus on the development of prediction models by applying the best suitable predictive algorithm tested, measured by its performance time and precision capability. These algorithms would be fed with the occupation data of the parking lot and context information, concluded on the first objective, which can turn out to be a challenge, since those types of data are hard to find and originate from a number of different data sources which makes it harder to acquire and integrate.

Lastly, the third objective consists on the development of a decision algorithm that takes into consideration various heuristics with an associated weight factor and returns the most suitable parking lot that takes into consideration current user and park conditions. This way an overall system like the one proposed can help not only people who do not know the city to find a possible parking space, but also residents who do, by telling them what is the best parking option.

1.3 Research Method

This research work followed an adaptation of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. This methodology is an open standard, robust and well-proven methodology, that provides a structured approach to planning a data mining project, as it can see in Figure 1.1 based on (Europe, 2018).

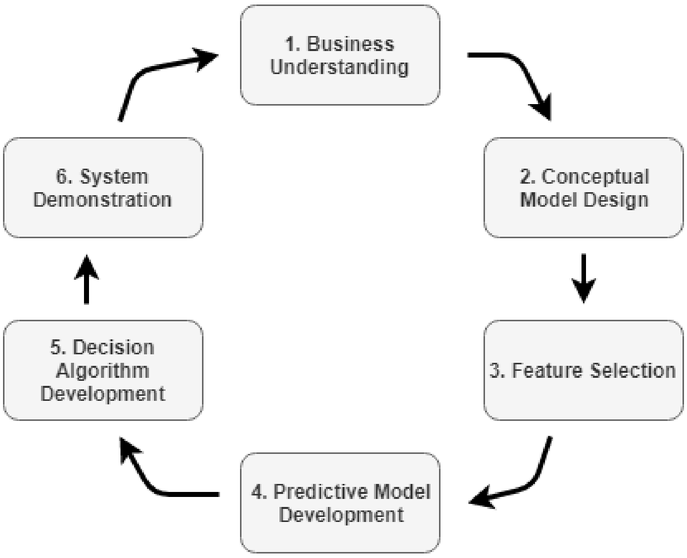

1. Business Understanding
2. Conceptual Model Design
3. Feature Selection
4. Predictive Model Development
5. Decision Algorithm Development
6. System Demonstration

FIGURE 1.1: Research Methodology Diagram.

In the first step there is the Business Understanding and can be further addressed in Chapter 1 and Chapter 2, where the needs of the business in question are analyzed. In the next step there is the Conceptual Model Design, seen in Chapter 3, where the proposed system is presented and is designed to be capable of dealing with availability problems. A brief description of each of the components developed is also addressed in this chapter. The next chapter, Chapter 4, the Feature Selection happens, where the description of each data source is made, as well as processing and treatment of the data. The data is then analyzed to try to understand which features are essential and helpful to the development of the system. Right after that, the Predictive Model Development phase occurs, seen in Chapter 5, where the data preparation and the testing of supervisioned algorithms is made, so the best predictive model can be created. After so, the Decision Algorithm Development starts, seen in Chapter 6, where this algorithm will be responsible of suggesting the most optimal parking lot for the driver to park. In the next step, Chapter 7, the System Demonstration is made so the overall functionality is shown.

1.4 Research Structure

This document is structured in 8 chapters, as it follows:

- In **Chapter 2** a literature review of models developed when dealing with the prediction of parking availability is made, but there was also a focus on surveying the most influential factors in the allocation of the parking lot and current available parking systems;
- In **Chapter 3** the conceptual model of the proposed system is presented and described for each component and how they work to deal with vehicle availability problems;
- **Chapter 4** is where the feature selection is performed on all of the different data types used in this study. The data sources are described in depth, processed and treated, with the aim of integrating the various types of data. There is also a focus in the analysis of all the different types of data, in order to understand if they have an influence on the occupation of the parking lot. Finally, there is the conclusion which data is the most important for the occupation rate prediction;
- In **Chapter 5** the development and testing of various predictive models is made by applying different algorithms to the data previously analyzed, with the aim of obtaining the best model to integrate in the system. Lastly, there is a focus on optimizing the predictive model in order to improve the results previously obtained;
- In **Chapter 6** the definition and development of the decision algorithm is made;
- **Chapter 7** focus on the demonstration of the proposed system;
- Lastly, **Chapter 8** shows the conclusions reached at the end of the work, as well as some future work to improve the overall usability and performance of the proposed system.

Chapter 2

Literature Review

In this chapter we focus on what type of systems are already in use and implemented, which features show to be more important and have a bigger impact on the parking availability and the type of models created to deal with this type of problem.

2.1 Systems for Parking Availability

Systems studied and available for the help of parking availability are based on various types of data sources used when dealing with this type of problem, based on real-time parking data, on user data and on historical data (Pflügler et al., 2016).

Systems based on real-time parking data can be equipped with Internet of Things (IoT) sensors. Real-time parking data is widely used and helpful in the prediction of park availability, as real-time parking information management could improve 10% of traffic in efficiency (Caicedo, 2010). Those sensors are able to output real-time information about the parking lot, like which parking spaces are free or not (Shin & Jun, 2014). The monitoring of each single parking space in parking lots is economically expensive, but in contrast it is quite feasible to monitor the flow of entering and exiting vehicles from a parking lot (Klappenecker et al., 2014), however the continuous IoT growth make it an increasing industry, resulting in lower costs for the sensors as well as more efficient and lower energy consumption (Zheng et al., 2015). Sensor information is produced at a large quantity and rate with huge heterogeneity, also known as Big Data, meaning that there is a necessity to apply efficient tools to analyze that data. A big disadvantage from this type of system is the cost of the equipment necessary for one single parking space, as well as the costs of maintenance (Rong et al., 2018). One big project involving parking sensors was *SFpark*, in San Francisco, using a total of 8,622 parking sensors, resulting on a total cost of 18 million US dollars (Richter et al., 2014).

Other systems based on real-time parking data use smart cameras capable of visual parking lot occupancy detection as we can see in (Amato et al., 2017), where this detection was based on a deep Convolutional Neural Network specifically designed for smart cameras. This systems can monitor more than one parking space at the same time, reducing the need to have sensors in every parking space and significantly reducing the costs of installing and maintaining the sensors. However, environmental conditions can greatly affect the performance of these systems, like shadows, light variations, partial occlusions of the image and weather conditions.

The data from systems based on user data are provided directly by users, also called crowdsourcing. This type of information can be reported by users passing by as a way to contribute to the system efficiency, introducing the concept of Gamification. Users are encouraged to give information about free parking spaces at that current moment, and if that information is correct, they could be rewarded with discounts on parking spaces and other benefits. Disadvantages of this type of system are the need to have a large user base, so that the data that is produced is enough to generate sufficient data to provide adequate parking information. Also, the information veracity is not ensured, meaning that there is a need to implement methods to make sure that accuracy is met, adding extra processing (Pflügler et al., 2016). Crowdsourcing can be executed by numerous ways, e.g. mobile sensors installed in recent vehicles or even mobile applications on phones are an alternative to sense complete city districts (Bock & Sester, 2016). Although, the crowd-sensing seems to be a good option from an economic perspective, its coverage is almost always incomplete and user dependent, since roads used very little will have essentially no coverage (Bock & Sester, 2016).

There are also systems based on historical available data, cost-effective and not as dependent on a user base. This type of system, if it has enough data, can cover cyclical variations over a year (e.g. seasons of the year, holidays period, and so on) which may prove to be important (Tilahun & Di Marzo Serugendo, 2017). Having access to historical data is really important when dealing with this type of problems, being easier to monitor and retrieve data from closed parking lots than on-street parking. Even in the off-street parks with no control admission, the information needs to be registered and such can be done with sensors, but the installation and maintenance of those types of hardware is very costly to cover a large area (Amato et al., 2017). Another easier solution is, instead of monitoring each single parking space, monitor the flow of entering and leaving in the parking lot (Klappenecker et al., 2014), this way the park monitor will always have the exact number of cars in the parking lot at a reasonable cost. However, this type of monitoring will not be able to give the exact position of a free parking space and can only be implemented on closed parking lots. In (E. I. Vlahogianni et al., 2016) six months of historical data was used, and in (Pflügler et al., 2016) only two months, revealing to be a short time to cover all possible outcomes and not showing the full impact of cyclic features like seasons of the year.

A critical problem with a data-driven prediction of a parking lot availability are the quality of the data used. Usually data from sensors shows a big amount of noise and variability and invalidation's (Bock et al., 2017), like missing values or incongruities, making the accuracy of the prediction worse, as it is harder to train a generalized model and harder to predict correct values (Richter et al., 2014).

Systems have been implemented to overcome the difficulty of finding a free parking space, namely the PGIS, that offer the driver information about the current state of the parking lot, like the free parking spaces. In (Grodi & Rios-gutierrez, 2016) a prototype of smart parking system using wireless sensor technology and networks, that captures the information about the parking space status and then shows it trough a mobile application or a website. Despite the fact these this systems reduce the economical and time costs, generally there is a high cost associated with the installation and maintenance of the sensors and do not take into consideration the time it takes to reach the parking lot.

As a way to deal with the parking problem and to cope with the limitations from PGIS, the study (H. Wang & He, 2011) developed a IPR system to optimize parking management with a reservation policy to balance the benefits of the service providers and the user requirements. (Shin & Jun, 2014) focus on creating a concept of smart parking guidance system with reservation proprieties (IPR) to assign the driver to the most appropriate parking facility considering various factors, parking cost, traffic congestion, distance to parking facility and walking distance to the destination. This system will monitor the parking lot status in real-time with the help of sensors and suggest the most appropriate parking facility based on the current status of parking lots and the information inputted by the driver. The user then has the opportunity to reserve the specific parking lot until they arrive and subsequently parking costs occur from the start of the reservation. These systems generally have a high cost for the user, that needs to reserve a parking space, with a cost associated to it, even before reaching the parking lot, where sometimes the user could have difficulties to reach the park due to traffic or weather conditions, increasing the total cost of reservation.

In Table 2.1 we can see a comparison of the different data sources used and their limitations and benefits.

Table 2.1: Studies comparison using different data sources.

Work	Data source	Benefits	Limitations
Temporal and Spatial Clustering for a Parking Prediction Service (Richter et al., 2014)	Real-time Sensors	Accurate	Expensive, Maintenance
Improving Parking Availability Maps using Information from Nearby Roads (Bock & Sester, 2016)	Crowdsourcing	Cheap, Gamification	User base, Veracity, Geographical coverage
A Real-Time Parking Prediction System for Smart Cities (E. I. Vlahogianni et al., 2016)	Historical	Cheap, Temporal coverage	Storage
Deep learning for decentralized parking lot occupancy detection (Amato et al., 2017)	Smart cameras	Monitor more than one space	Environmental conditions

2.2 Feature Selection for Parking Availability Forecast

There are some factors that can influence the search for a free parking space. Weather information is one of the features that reveals to be important when evaluating the parking occupancy, like rain intensity, temperature and wind strength (Lijbers, 2016). The author in (Greengard, 2015) says that the weather data has a real impact on the traffic behavior, more specifically rainfall and temperature, making them important factors for the parking prediction. Bad weather conditions could lead to lower traffic flow than expected, in the work of (Yang, Liu, & Wang, 2003) the weather information has a big impact on the traffic data, namely the traffic flow intensity, but parking occupancy would just be affected in shopping malls, iconic locations, and other, not on parks close to apartments and offices (Rong et al., 2018).

The location of the parking lot is also an important influencing factor on the parking occupancy (Mathur et al., 2010). In the (Tiedemann, Vögele, Krell, Metzen, & Kirchner, 2015) the authors concluded that the typical occupancy behaviour on a parking lot depends mainly on the location (e.g. residential vs commercial area) and the day of the week, and it should be expected that certain external factors, like holidays, winter weather, street works and events lead to changes in previously identified patterns. In (Rong et al., 2018) each parking lot is categorized into seven categories of the parking lot, apartment, office, mall, food, hospital, park and entertainment. If the parking lot is close to shopping centers or supper markets, it is categorized as a mall parking lot, and so on. The idea is that shopping malls will have a different availability from 8 AM to 5 PM, than a park from an office building, and models created for a type of category can be replicated to other parking lots inside the same category.

One important factor that may over-saturate the parking lot are the events on the surroundings of the location of the park (Xiao et al., 2018). If the parking lot is in the proximity of some type of shopping mall or close to an important public highway, or even if events happen regularly around the parking lot, like soccer games and concerts, those can cause a significant increase in the amount of traffic, consequently increasing the demand for free parking spaces (Ionita et al., 2018). Events like concerts or soccer matches tend to be a impactful factor on the parking occupancy (Yang et al., 2003), as those occurrences lead to an increase of traffic volume. If rich historical data can be implemented and information regarding the events is known in advance, the prediction can better adjust itself to take into account those special occasions. So, when predicting parking availability, factors like spatial and temporal have varying importance (Rong et al., 2018), so first there must be an evaluation on which features should really be used, since data like traffic and events are harder to get and the effort to integrate that information is increasingly higher (Pflügler et al., 2016).

The period of the day and time of year are also important (Zheng et al., 2015), as holidays, weekdays and hour of the day could have direct impact on park occupancy. The time of day seems to be a very important factor also according to

(Z. Wang et al., 2007), where holidays, weekdays and time of the day could lead to different parking situations. Once again (Greengard, 2015) agrees with this conclusion, stating that the traffic volume varies very much depending if it is a holiday and if we are in a vacation period. The information of the time of the day could lead to a different conclusion, namely the time of the day, the day itself, month, year, holidays and vacation periods affecting, once again, the traffic and park occupancy (Pflügler et al., 2016). For (Pullola, Atrey, & Saddik, 2007) the parking lot occupancy varies depending on various criteria, such as the day and time and traffic situation. In the work (Chen et al., 2013) the authors concluded that taking into account exogenous variables, like daily/weekly/seasonal patterns, or the effect of the weather, greatly increases the prediction accuracy over other models not using them.

Holiday features reveal to be really important, as parking availability is really different between a normal day and an holiday, showing bigger parking occupancy in parking lots close to apartments and shopping malls, and more quiet in office parks (Rong et al., 2018). In the case of (Chen et al., 2013) the month of December was categorized as Christmas Shopping Season (1 December to 31 December), where there is a strong impact on the demand of parking lots, the demand in parking lots is bigger during weekends than in the weekdays, since many people take those opportunities to go shopping inside the city center. This demand is so high in this time of the year that it leads to fully occupied car parking on weekends, regardless of rain or fog.

Analyzing the park features can be relevant to determine the drivers' behavior towards that parking lot, like how close are they to the final destination of the user, the prices charged by the car park authority (Giuffrè et al., 2012), the time it takes to have a free parking space, the duration and distance of the path from the current location to the respective parking lot. For (Shin & Jun, 2014) the parking cost and estimated queuing time outside the parking lot are important factors to be taken into consideration, which can be used to evaluate the effectiveness of the parking guidance. The estimated queuing time can be calculated with the help of mathematical programs, as we can see in (Thompson, Takada, & Kobayakawa, 2001) where one of the programs objectives was to give minimal queue length.

As we can see throughout all features we can conclude that traffic information is one of the most important factors when predicting the availability of a parking space, as it directly influences the parking occupancy (E. I. Vlahogianni et al., 2016).

Table 2.2 shows the resume of the features that influence the parking availability.

Table 2.2: Studies results about the features that influence parking availability.

Work	Features
Predicting Parking Lot Occupancy Using Prediction Instrument Development for Complex Domains (Lijbers, 2016)	Weather information
Parknet: Drive-by sensing of road-side parking statistics (Mathur et al., 2010)	Location
Du-parking: Spatio-temporal big data tells you realtime parking availability (Rong et al., 2018)	Park category, Holidays
Predicting the Availability of Parking Spaces with Publicly Available Data (Pflügler et al., 2016)	Time features (month, day, hours, etc)
Novel Architecture of Parking Management for Smart Cities (Giuffrè et al., 2012)	Parking lot characteristics
How likely am I to find parking? – A practical model-based framework for predicting parking availability (Xiao et al., 2018)	Events
A Real-Time Parking Prediction System for Smart Cities (E. I. Vlahogianni et al., 2016)	Traffic

2.3 Forecasting Parking Availability Techniques

Various types of techniques have been applied to predict and inform the user about the future parking availability for on-street parking spaces and for closed parking lots.

In (Liu, Lu, Zou, & Li, 2006) the authors presented a time series capable of predicting the daily parking demands in an intelligent transport system, while also concluding that predicting parking availability on specific hours is not effective based solely on parking data. The choice of the parking lot depends on user preferences and parking space availability, parking fees and the distances between parking facilities and final destinations.

A good concept is the development of a system that can integrate with the current GPS based navigation systems from vehicles, and that is what is proposed in (Pullola et al., 2007), where the system gives information to the user about which parking lot in the surrounding area of the users destination has the most probability of having a free parking space. The availability of the parking lot, at a corresponding time, is modeled using a non-homogeneous Poisson process. This model takes into consideration the past availability data under different contexts for the time interval, the current contextual information for the time interval and at the availability at the current time. However, this integration is quite difficult as current vehicle systems have limitations in terms of storage. So, in (Richter et al., 2014) it is proposed a back-end based approach to learn models of parking availability with historic data in order to save those models to the on-board navigation systems of the vehicle, as those models are really compacted. Compacted models, generated with the help of clustering, can reduce the storage space need up to 99%, maintaining the prediction accuracy of around 70% (Richter et al., 2014).

In (Klappenecker et al., 2014), the authors believe that the future availability of the parking lot from malls and airports, having a way of controlling the admission on the park, are not affected by the past occupancy, but due to present. Following that idea, they introduce a model based on a homogeneous continuous-time Markov chain, representing the changes of the parking spots over time.

The focal point in the study of (Chen et al., 2013) is the development of a predictive model to predict the availability of parking spaces or bikes from share bicycle scheme for short-term (5 minutes ahead), medium-term (1 hour ahead) and long-term (24 hours ahead), using a Generalized Additive Model that takes into account exogenous variables, like weather conditions, time of day, if it is a weekday, weekend or holiday and the year. The weather conditions are only taken into consideration for the short-term conditions, and not for the medium and long term predictions. When compared with existing methods, there was a significantly improvement on the performance for the three time periods, resulting on a Root Mean Square Error (RMSE) of 15.8% with a 2.5% standard deviation.

In (E. Vlahogianni, Kepaptsoglou, Tsetsos, & Karlaftis, 2014) models are developed for predicting the parking occupancy and the time period a parking space remains free. Those models are developed as a real-time series occupancy prediction scheme based on artificial neural networks, with the help of past information about parking, as well as other variables, like traffic volume, weekends, weekdays and time period (peak, off-peak, morning, evening). The models are developed from 1 minute ahead prediction to 60 minutes ahead predictions, resulting in good outcomes, especially for 60 minutes predictions using a Multilayer Perceptrons of 8 hidden layers and a historical information of about 5 minutes earlier when comparing with a naive prediction. Later, the authors tackle this problem once again in (E. I. Vlahogianni et al., 2016), where data was obtained wirelessly from a IoT sensor network available in the "smart" city of Santander, Spain, giving the current status of the parking space (free/occupied). In this work the parking efficiency can be defined by two metrics, the average time duration that a slot is free over a certain time period and the percentage of parking slots occupied during a predefined time period. A module is proposed as a real-time time series occupation prediction based on recurrent artificial neural networks in order to predict an overall occupancy of parking spaces while using past information and results show an increase of accuracy when compared to a naive prediction, and when the predictive horizon increases, the better are the results when comparing to the naive technique, with a mean absolute percentage error of less than 3.6% for a 15 minute prediction for various parking regions.

The work developed by (Zheng et al., 2015) focuses on the analysis of three non-parametric models with three different time-series feature sets, one with the time and day of the week features, the second with N measurements before the time to predict and lastly, a combination of the two feature sets, all of them output the occupancy rate (between 0 and 1) at the specific time t. For modelling the occupancy rate and for the prediction, the authors used Regression Tree, Support Vector Regression (SVR) and Neural Networks (NN). The predictions were made for 15 minutes, 1 step ahead, and for k * 15 minutes, k being the number of steps ahead, but for predictions higher than 15 minutes SVR was not applied, due to the long computation time needed. The authors were able to conclude that the feature set showing better results was the one that joins the features of the previous feature sets and that the Regression Tree, which is the least computational intensive algorithm when compared with the other two techniques, performs best for parking availability prediction when comparing with the NN and SVR.

In (Richter et al., 2014) the proposed work focused on reducing the number of models necessary to predicting parking space availability on various road segments, by using different clustering of the historical data from road segments composed by 5 minute time slots over a 24 hour period, allowing also to predict parking availability in a long-term, by applying Markov Chains. This is possible by averaging the captured values for each time slot with the clustering techniques, this way the aggregation eliminates the distinct park peaks due to outlier events, like soccer games, but identifies the general parking trend for each road segment. The work concluded that a seven-day model had the best predictions (78% accuracy), while also having the highest storage requirements, in contrast the temporal clustering resulted in the worst results with an average of 66% of accuracy and the spatial clustering resulted in an average of 68% of the accuracy, meaning that both solutions can decrease the need of storage up to 99% but cannot obtain as good values as the seven-day model. The work of (Bock & Sester, 2016) also shows an evaluation of spatial similarities in parking availability based on the parking sensors in the city of San Francisco, revealing that relevant spatial similarities in parking availability can be seen in distances bellow the 100 meters mark. When comparing time similarities with the spatial similarities, parking availability rates are lower, but can still be useful for nearby roads. Some spatial interpolation methods were tested, but the inverse distance weighting method had the best results. In this work, the authors conclude that a combination of both spatial and time information from nearby roads is promising when predicting parking availability.

One helpful solution to predict more accurately the free parking spaces in a location, is dividing the area into smaller and equal cells, divided by street segments (E. I. Vlahogianni et al., 2016) or divided by a geographic grid (Rong et al., 2018), as different independent variables are significant in different parking regions when checking the free parking space duration.

The idea behind (Ionita et al., 2018) study is creating a parking demand profile using K-Means to cluster areas, as a way to scale prediction systems where there is no parking data. This solution reveals to be a good option to reduce the implementation costs of this type of model in other areas of the city, since the sensors installation and maintenance has a very high cost. With the help of machine learning techniques, more specifically decision trees over random forests, support vector machine, multilayer perceptrons and extreme gradient boosting, the authors were able to predict the parking occupancy for every cluster with the help of the average price and average capacity per block, and evaluate each of the outputs with the help of the RMSE metric. The extreme gradient boosting showed the best results when applied to data clustered in 8 or 16 clusters and that clustering the city into smaller areas produces better occupancy estimations than entire city models.

For the (Tiedemann et al., 2015), the different occupancy behaviour can differ a lot and new factors might come up often and quickly (e.g., street works and events), for those cases historic data could improve the prediction quality a lot.

In order to get the best prediction, the model not only uses the time of day and the day of the week, it also uses the current occupancy situation at the time of the query, this way the most recent measurements can be used to adapt to very recent changes in the parking occupancy.

The main focus in (Pflügler et al., 2016) is the development of a prototypical system for the prediction of the parking situation using only publicly available data, thus reducing the costs, such as the need to implement sensors in parks, while also identifying important data sources to help in the prediction process. The authors implemented a system based on a NN reaching an average square error of 0.16321. The study also concluded that weekday, location, temperature and time of the day improve and enhance the prediction accuracy, while traffic, holidays, events and rainfall has a secondary relevance.

In (Bock et al., 2017), after applying the extraction of trends, a generic regression model is trained with the extracted trends using once again SVR, to produce the availability prediction. This model shows to be more accurate than the baseline, created using the SVR model with the raw data without any treatment. With the 2-step approach there is also another benefit, namely the reduction of the parking availability dataset size, as the trend data reduces the size of the dataset by about 60%, while maintaining valuable information for the PGIS.

Using data that covered a total period of 3 months of parking occupancy recorded by on-street parking meters of two specific zones in the central area of Lisbon, that are surrounded by residences, universities, commerce and event venues, (Ramos Silva, 2017) developed a classifier that indicates the parking situation from a multiple range of classifications, namely vacant, almost full and full. During this work various algorithms were tested, namely J48, Random Forest, REPTree and Multi-Layer Perceptron, with the help of some contextual attributes like, the class hour, weather conditions, temperature, precipitation, holidays, vacations, week number in the month, begin month, end month, special events and outliers. Results show that Random Forest had more consistent results.

For the development of the application ”Du-Parking” (a deep learning based approach) (Rong et al., 2018), the authors estimated a real-time parking availability throughout the city only using historical parking data and a variety of parking datasets, like weather, events, map mobility trace data, holiday, POI-related features and navigation data. Three techniques are used to evaluate the precision and recall of the model for the prediction of the parking availability, namely linear interpolation, gradient boosting decision trees and deep neural network, where the results show that gradient boosting decision trees outperforms linear interpolation, and the deep neural network increases the accuracy when comparing to the gradient boosting decision tree, also concluding that the temporal information is more beneficial for this problem.

In (Rajabioun, Foster, & Ioannou, 2013) the authors presented a new parking guiding system to assist the user to find the most suitable parking space based on the user’s preferences. This work also focused on developing a prediction algorithm to forecast the number of available parking spaces in both on-street and off-street, in order to increase the preciseness of the guidance. Information like the availability, price, parking rules, location, type of parking and others was made available via web.

To build a robust model for the prediction of parking availability there is a need for large amounts of data, resulting on large processing times, resulting on large economic costs for the system when predicting parking availability in multiple areas of the city, so in (Zhang & Li, 2018) the authors proposed a deep learning based parking prediction system where all components of the system are based on cloud platform. For the prediction of the parking availability, a Long-short term memory network is used while taking into account multiple factors such as time of day, weather and holiday. Long-short term memory shows better accuracy values when comparing with the other tested technique, BP Neural Network, where the RMSE showed a value of 5.42 for the Long-short term memory network and 13.87 for the BP Neural Network, but in contrast the mean prediction time is 18.03 seconds for the first technique and 9.65 second for the BP Neural Network case.

In the (Xiao et al., 2018), the authors decided to estimate the parameters that feed the predictive model for the real-world applications. Those parameters will give the model the idea of the arrival rate and departure rate through the parking time, helping the predictive model. Classification of the day-to-day patterns were made, dividing the day data into workdays or holidays, as well as regular workdays or high demand days or holidays. The prediction of the future occupancy can be engaged by using or not using real-time updates of the occupancy, but it would be beneficial to monitor real-time data as a way to make adjustments for when special events occur or other unforeseen reasons that can not be reflected in historical data set.

Work performed in (Stolfi, Alba, & Yao, 2019) takes an approach of developing a system capable of collecting public data on car park occupancy values and display them in a web service, while also storing this information in order to be able to predict the car occupancy rate in future weeks. To obtain the best results, the authors decided to test the accuracy and complexity of various algorithms, namely Polynomial Fitting, Fourier Series, K-Means, KM-Polynomial, Shift and Phase and Time Series, predictors used in previous works (Alba, Chicano, & Luque, 2017), with datasets from 4 different locations. Time Series showed to be the most accurate algorithm for all the 4 locations, although it requires a larger amount of data to represent each car park and weekday.

Finally, the goal of the work (Ziat, Leroy, Baskiotis, & Denoyer, 2016) is to provide the users with a way to optimize their travel plans by giving them traffic and parking occupancy predictions. A representation learning model for time series forecasting is proposed and compared to more traditional techniques, like mean, vectorial auto-regressive, neural network and autoregressive integrated moving average. Results show that considering time series outperforms the classic techniques, for all the prediction horizons.

Table 2.3 represents the various techniques used in some studies to predict the parking availability.

Table 2.3: Studies results about the different prediction techniques used for parking availability.

Work	Technique	Results
Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots (Chen et al., 2013)	Generalized Additive Model	RMSE = 15.8% with a 2.5% standard deviation
A Real-Time Parking Prediction System for Smart Cities (E. I. Vlahogianni et al., 2016)	Multilayer Perceptrons	Mean absolute error percentage (MAEP) <3.6%
Parking availability prediction for sensor-enabled car parks in smart cities (Zheng et al., 2015)	Regression Tree	MAEP = 5.7%
Temporal and Spatial Clustering for a Parking Prediction Service. (Richter et al., 2014)	Markov Chain	Accuracy = 78%
Where to Park?: Predicting Free Parking Spots in Unmonitored City Areas (Ionita et al., 2018)	Extreme Gradient Boosting	RMSE = 14.52% to 22.93%
Predicting the Availability of Parking Spaces with Publicly Available Data (Pflügler et al., 2016)	NN	Average Square Error = 0.16321
Predicting Space Occupancy for Street Paid Parking (Ramos Silva, 2017)	Random Forest	Precision = 70% to 75%
Du-parking: Spatio-temporal big data tells you realtime parking availability (Rong et al., 2018)	Deep Neural Network	Precision = 84.47%

Chapter 3

Conceptual Model of Park Aid App

This chapter focus on showing the conceptual model of the system being developed to deal with various parking availability problems inside a big city, as it can see in Figure 3.1. In this case the principal focus is to guide a driver to a parking lot with an available parking space. The system is composed by various components, namely the Final Dataset, Predictive Model and the Decision Algorithm. There was also the development of an interface where the proposed solution would be integrated, so the users could interact with it. The interface created is an android application, named "Park Aid", where the user can interact with the system and get information about the parking lots, namely the current occupation rate, the price per hour, the working period, and more.

This conceptual model follows a similar architecture of what is described in (Alface, Ferreira, & Pereira, 2019), with only a few components additions, but framed with the parking availability paradigm instead of electric vehicle charging.

In the following subsections each component of the proposed solution being created is being presented.

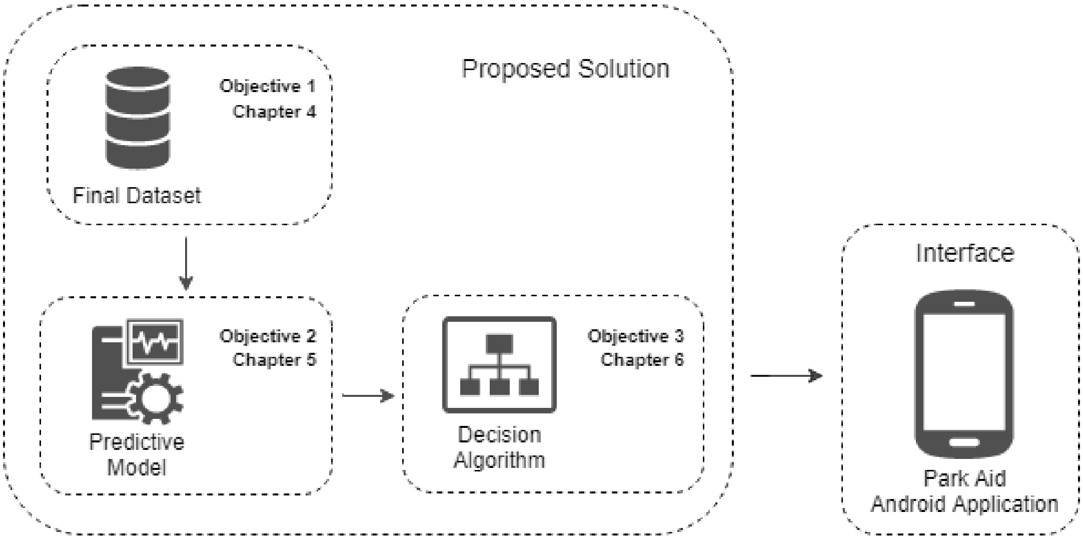


FIGURE 3.1: Conceptual Model Diagram

3.1 Final Dataset

In Chapter 4 three different data sources are analyzed and a feature selection is made to create the Final Dataset. This dataset is going to be used in the construction of the predictive model and so maintaining the most helpful features is important since these features increase the prediction accuracy of the parking availability, being this the first objective of this study. Also, the removal of useless features contributes to the reduction of the computational time required and reduces possible prediction errors.

The first data source being used is the parking occupancy data from three parking lots in Lisbon situated around the Marquês de Pombal area, from 1 October 2018 to 31 January 2019. The second data source is the weather data of Lisbon, with hourly historical weather data, and the third data source is traffic data of Lisbon, able to provide traffic information in the surroundings of each parking lot.

Historical data is very cost-effective and does not depend on the user, since it is possible to understand cyclical patterns (Tilahun & Di Marzo Serugendo, 2017). Although the current data is not sufficient to cover the annual pattern, historical data is really important when dealing with the prediction of the park occupancy, giving us the advantage of generalized trends over time periods (Richter et al., 2014). To have this coverage, the information of the occupation values in the parks is saved by monitoring the number of cars entering and exiting the parking lot, allowing you to have the precise number of cars in the parking lot without incurring a large associated cost like when monitoring every single space (Klappenecker et al., 2014), as is the case with parking lots using sensors (Amato et al., 2017), but in contrast measuring flow on the entrances and exits of the parking lot is not capable of giving the exact position of a free parking space and can only be implemented on closed parking lots, as is our case.

3.2 Predictive Model

Next component is the Predictive Model, being exploited in Chapter 5 and corresponding to the second objective, that gives the availability rate of a parking lot at an interval time. This Predictive Model is created by applying a supervisioned predictive algorithm from the python H2O library (H2O.ai, 2019) to the Final Dataset obtained after the feature selection performed in Chapter 4. The predictive model is built in a local server and then exported to the ”Park Aid” application, having certain effects on the final solution, like increasing the total size of the application and the battery drainage. The algorithm chosen is evaluated in Chapter 5, as various algorithms are tested and examined to decide which one is the most desirable, while taking into consideration the accuracy levels and the performance, since the execution time is important inside an android application. The developed approach can be applied to any parking lot with similar characteristics, by introducing the necessary features. Further details from the predictive model and the data used in the creation of the model are further explained in Chapter 5.

3.3 Decision Algorithm

The last component is the Decision Algorithm which meets the third objective, being developed in Chapter 6. This algorithm will be responsible for suggesting the most suitable parking lot for the driver. The decision algorithm will take into account various heuristics to be able to output which parking lot best fits the current driver, heuristics like the current driver's position and the parking lot characteristics, as well as the parking lot occupancy obtained through the developed predictive model in Chapter 5, are taken into consideration and are necessary for the proper functioning of the decision algorithm.

3.4 Park Aid Android Application

”Park Aid” is the interface responsible to give the user a friendly and intuitive way to interact with the proposed solution. The main idea of the App is to contain information about the various parking lots so that each one has a predictive model associated that provides the occupancy rate of the parking lot at any given time, as well as guiding the user to the best parking lot option suggested by the decision algorithm.

To improve and make it easier for drivers to find an available parking space, two options named "Navigate" and "Check Map" have been created. These options present information and the location of the parking lots, as well as the parking lot characteristics in order to enable users to plan and choose the right route. This options are also able to plot the route from the users current position to the chosen destination parking lot, so that users who are not familiar with the city have help in finding the final destination.

The first step when entering the application is the login, where the user needs to put his credentials to log in the application, and in case it is the first time logging in, the user can register a new account by providing the email, username and a password. After filling the login credentials and entering the app, the user is greeted with a two option menu, shown by the Figure 3.2, being explained in the next paragraphs.

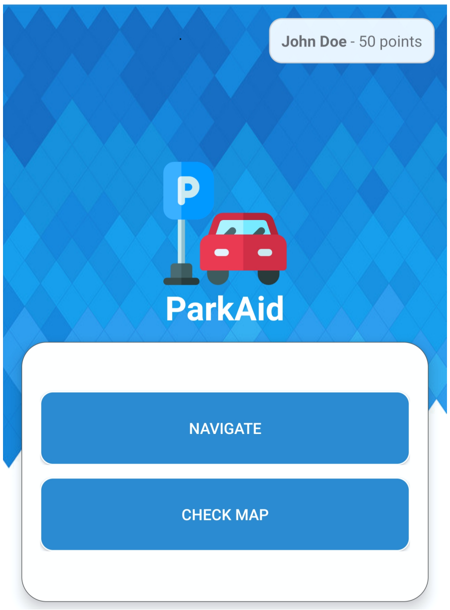


FIGURE 3.2: Park Aid application menu.

For the first option of the application, “Navigate”, the driver needs to provide the final destination to travel to, as well as the maximum walking distance willingly to walk, as a way to restrict the search zone. That is, the recommendation of the most optimal parking lot will be carried out within the search radius provided, centred on the final destination. This recommendation is made by the decision algorithm that is developed further in the work, namely in Chapter 6. If no parking lot is found, the closest parking lot will be recommended to the driver.

The second option shown in the menu is “Check Map” and when clicked a map focused in the current user location appears. This map shows markers that represent the location of all of the parking lots, shown in the Figure 3.3, and when a marker is clicked a popup information board appears with information of the parking lot, as it can see in Figure 3.4. The information shown is the probability of occupation of the parking lot, the name and description of the parking lot, the address, the working period, the parking fees and a button (”Go to location”) that creates a route to the respective parking lot, having as the starting point the current user location. In this case, the maximum meters the user is willing to walk from the final destination to the parking lot is not taken into account, since this information is not supplied, but information about the percentage of occupation in the park at the end of the route is shown, as well as the duration it would take to reach the park.

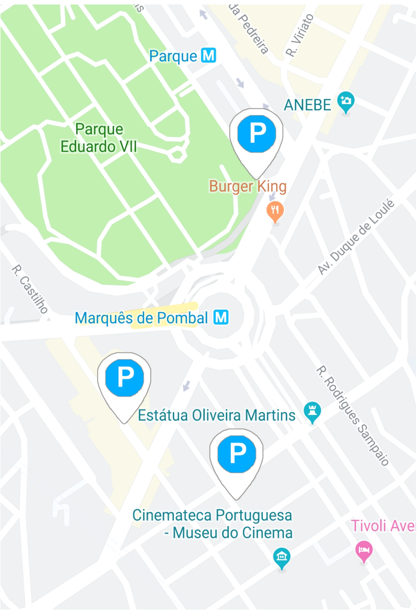
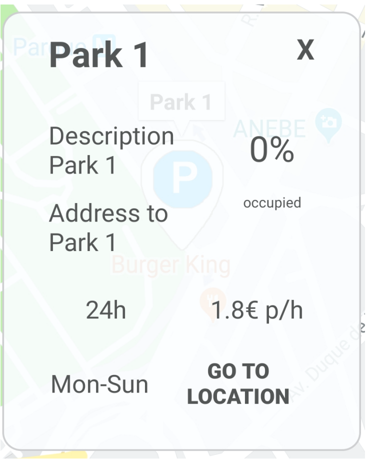


FIGURE 3.3: Parking lots location on the map.


Park 1

X

Description Park 1

0%

occupied

Address to Park 1

24h

1.8€ p/h

Mon-Sun

GO TO LOCATION

FIGURE 3.4: Pop-up information board on a parking lot.

Chapter 4

Feature Selection

The main focus of this chapter is to analyze the data to have better insight about the vehicle influx in parking lots and make a feature selection of the most influential features on the parking occupation in Lisbon, more specifically in the Marquês de Pombal area. This data is analyzed for the period of 1 October, 2018 to 31 January, 2019, making it a total 4 months of data. In this study three parking lots were approached and even though the data does not have an extensive size to see the annual pattern of the parking lots occupancy, it intercepts a key moment for park affluence, namely the Christmas period. This period allow us to analyze and perceive how the parking occupancy changes during festive periods and how holidays can impact the parking lots occupation (Chen et al., 2013).

This chapter is really important in order to reduce the number of data and its complexity to be used in the production of the predictive model, to improve the results of the models and to reduce the necessary size on the mobile phone. In this chapter the analytic process, seen in Figure 4.1, is followed.

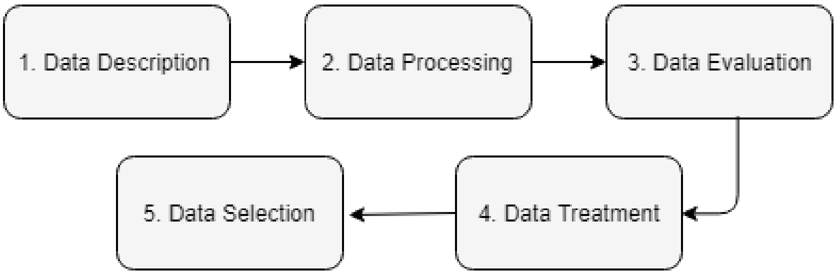


FIGURE 4.1: Feature Selection Process

As it was said earlier, the process used to run the feature selection can be seen in Figure 4.1, where in the first step, Data Description, the data used in this work is described in depth. In the second step, Data Processing, some data processing is performed in order to add new information to the data. In the next step, Data Evaluation, the data is evaluated to better understand the trends of occupation, as well as identifying outliers cases and justifying them. In this step two techniques were also used to evaluate the influence of each feature on the occupation rate of each parking lot measured by hour. The fourth step, Data Treatment, is where techniques to treat invalid measurements are presented and applied, and in the last step, Data Selection, the selection of the relevant and useful data is made to be used in the rest of the study.

4.1 Data Description

In this section, the identification and description of the used historical data from multiple sources is made. It is also important to know that the traffic and parking lot data were obtained via a non-disclosure agreement, so the source of the data can not be revealed.

4.1.1 Parking lot data

The main data used is from the parking lots, and in this case there are three datasets in the Comma-Separated Values (CSV) format, one for each parking lot being studied, that were named as Park 1, Park 2 and Park 3. The data was gathered every hour between the period of 1st October, 2018 to 31st January, 2019. The first park is the biggest one, with a total of 1081 parking spaces, the Park 2 has 336 parking spaces and the final park, Park 3, has 154 available places.

All of the parking lots are underground parks, having parking spaces for people with reduced mobility and in the case of Park 1 it also has places to charge electric vehicles, which nowadays can be a differentiating factor when users are choosing which park to go to or agree on a covenant. The parks also have extra services like WCs, car washes, enhanced security with CCTV and vending machines, that may be decisive for users adherence to parking lots. Also, all of the three parking lots are open 24 hours a day and for the seven days of the week.

Another important factor is that the parking lots are located on a area which is surrounded by office buildings, so the parking lots were categorized as office parking lots which may prove important in terms of their affluence and time periods (Rong et al., 2018). The location of the parking lots also turns out to be quite important at the moment of decision by the users (Giuffrè et al., 2012), since a good location can define the use of a park. A decisive characteristic for the parking lot occupancy is the association with a cost per hour (Shin & Jun, 2014) and all of the parking lots studied have hourly prices. To enter the park, the driver needs to pick up a ticket that has to be provided when leaving the park and paid for by the total number of hours that the vehicle remained within the park. The cheapest parking lot is Park 1, with only a 1.80€ price per hour, while the Park 2 is 2.15€ and the Park 3 is the most expensive for a total of 2.30€. In Table 4.1 the price table for each parking lot can be seen.

TABLE 4.1: Parking prices for the 3 parking lots being analyzed.

Parking Lot	First 15 minutes	Hourly	Daily Maximum
Park 1	0.60€	1.80€	13.00€
Park 2	0.75€	2.15€	15.00€
Park 3	0.55€	2.30€	19.00€

The data came in the format of CSV composed by four columns, date, hour, rotation and covenant. The hour and date column represent the hour and date the measurement of the number of the vehicles were made, respectively, so they were combined, resulting on a *datetime* column.

The rotation column represents how many rotation vehicles are inside the parking lot for that measurement. Rotation vehicles are the type of vehicles that enter and leave the parking lot without any commitment, besides having to pay the ticket for the total number of hours spent in the park.

The covenant column represents the number of vehicles that have some agreement with the parking lot entity, being able to enter and leave the parking lot whenever they want for the time period they paid. Every parking lot has a maximum number of covenants and a spot is always reserved for the vehicle with the covenant. There are multiple covenant types in the parking lots being studied, namely the 24h covenants, daytime covenants and night time covenants, but since there is no way to identify what type of covenants it corresponds from the data provided, all covenants were considered as 24h covenants, where this type of covenant gives the user unlimited entry and exit from the park during the month in which he made the advance payment.

Combining the two columns gives us the total number of vehicles inside the parking lot for the respective time.

4.1.2 Weather data

The second type of data used to perform this analysis was the weather data in Lisbon acquired from the OpenWeatherMap History Bulk Application Programming Interface (API) (OpenWeatherData, 2019), from 1 October 2012 to 14 March, 2019. The acquired data came in a CSV file format with a total of 50947 rows, where each row of the dataset represents a measurement done for a respective date and time, collected in intervals of 1 hour as the weather conditions typically do not change much during short time horizons (Chen et al., 2013). The CSV file came with a total of 25 columns where 6 were used, namely the ones in Table 4.2.

TABLE 4.2: Columns used from the weather data.

column	description
datetime	date and time the measurement was made
temp	current temperature in Kelvin
humidity	humidity in %
wind_speed	wind speed measured in meter per second
weather_main	group of weather parameters (Rain, Snow, Fog, etc.)
weather_description	weather condition within the group of weather parameters

The *weather_main* feature represents the weather condition within the following categorizations on our data: clear, clouds, drizzle, fog, mist and rain. The *weather_description* gives some more information within the *weather_main* condition having the possible results shown in Appendix A.

4.1.3 Traffic data

The other type of data used for the enrichment of the analysis was the traffic data. This data gives us information about the traffic state on certain roads and the amount of time the vehicles need to go through a road in a certain moment. As it has been concluded in the literature review, traffic information is one of the most important factors when predicting the availability of a parking space, as it directly influences the parking occupancy (E. I. Vlahogianni et al., 2016).

The traffic data was gathered for the surroundings of the parking lot since those areas in Lisbon are heavily influenced by traffic. The data was obtained for the same period as the parking lot data, more specifically from 1 October, 2018 to 31 January, 2019, and for the surroundings of each parking lot. This data came in the format of a JavaScript Object Notation (JSON), with various components, in particular the following ones in Table 4.3.

TABLE 4.3: Columns used from the traffic data.

column	description
datetime	the date and time the measurement was made
segment_id	uniquely identifies the segment of road
average_travel_time	the average time it took the vehicles to pass through the segment of the road in seconds

The chosen roads for each parking lot can be seen in the following bullet points:

- Park 1 roads: Praça Marquês de Pombal, Túnel do Marquês de Pombal, Rua Braamcamp, Avenida da Liberdade, Avenida Duque de Loulé, Rua Joaquim António de Aguiar, Avenida António Augusto de Aguiar, Alameda Edgar Cardoso and Avenida Fontes Pereira de Melo;
- Park 2 roads: Rua Alexandre Herculano, Rua Braamcamp, Rua Castilho, Rua Mouzinho da Silveira, Rua Duque de Palmela, Praça Marquês de Pombal and Túnel do Marquês de Pombal;
- Park 3 roads: Rua Castilho, Rua Braamcamp, Avenida Engenheiro Duarte Pacheco, Praça Marquês de Pombal, Túnel do Marquês de Pombal and Rua Joaquim António de Aguiar.

4.2 Data Processing

In this subsection the transformation and manipulation was performed on the previously identified datasets. For this, several tools were used together, namely Microsoft Excel and the programming language of Python, using libraries such as Pandas and Numpy.

This process started by initially merging all the previously referenced datasets by the *datetime* column all of them have, making it easier to analyze all data and take conclusions. The merge process resulted in three datasets, one for each parking lot, composed by the park, weather and traffic data.

The first treatment performed was transforming some columns to a unit that is clearer and simpler to analyze, such as converting the column *temp* from the Kelvin unit to Celsius and the *wind_speed* column converted from meters per second to kilometers per hour.

From the traffic data, the *average_travel_time* column from the previously selected roads for each parking lot were used, resulting in a new column per road representing the value of *average_travel_time* from that road for the respective date and time the measurement was made. This means that for the Park 1 9 new columns were added, for Park 2 7 columns were added and finally 6 new columns were added to Park 3.

In the next step new columns were created, namely a *total_occupation* which consisted in the sum of the *covenant* and *rotation* columns. A *occupation_rate* column was also added with information about the rate of the occupation in the parking lot, instead of the true and continuous value, being calculated by the total number of vehicles inside the park divided by the total available places in each parking lot, just like in (Alface et al., 2019) where this transformation resulted in better accuracy values and also allows for better and more intuitive analysis. This means that if the Park 1 has 850 vehicles inside, it results in a 78.6% occupation rate.

After that, new information derived from the *datetime* were added, namely a year column, month column, day column and hour column informing the year, month, day and hour of the measure, respectively. With the help of the *datetime* column, a column named *dayofweek* was also added, giving information of the current day of the week, where the value 0 represents Monday, 1 represents Tuesday, and so on, until Sunday that has value 6. The *flag_weekend* that identifies if the current day of the measurement is on a weekend was also added, having value 1 if so, and value 0 if it is a workday.

Various flag columns were added, the first one was the *flag_holiday* that identifies if the current day represents an holiday or not. For that all holidays that occur during the period being studied were collected, as holidays are also important and have a direct impact on park occupancy (Zheng et al., 2015). For the data used in this study, a total of 10 holidays were selected with different levels of importance, meaning that for the same time span of the parking lot data not all of the holidays have the same importance, i.e. Christmas Day has a bigger impact then the "Dias de Reis" (Kings' Day). An importance value between the range 0 to 2 was given to the holidays, where 0 is a normal day, as those do not represent any type of public holiday, value 1 represents a festive day, not representing an officially public holiday, and 2 a very important and official public holiday. Results about the holidays and their respective value can be seen in Table 4.4.

Table 4.4: Public holidays used for the analysis.

Date	Public Holiday	Importance
05/10/2018	Implementação da República	2
01/11/2018	Dia de todos os Santos	2
01/12/2018	Restauração da Independência	2
08/12/2018	Dia da Imaculada Conceição	2
25/12/2018	Natal	2
26/12/2018	Boxing Day	1
31/12/2018	Réveillon	1
01/01/2019	Dia de Ano-Novo	2
06/01/2019	Dia de Reis	1

A vacation period between 22 December, 2018 to 2 January, 2019 was defined where usually people take Christmas and New Year's Eve vacation to celebrate this period. So, a new column identifying this period was added with the name *flag_vacationperiod*. This column has value 1 for every measurement made inside the interval and 0 for every other case.

The last flags added represent the current weather condition based in the *weather_main* column. The first one was the *flag_fog*, identifying measurements where the atmosphere had any type of fog, this column would have value 1 if the *weather_main* column had one of the following results: fog and mist. The *flag_rain* was also added giving information if during the measurement it was raining, having value 1 if *weather_main* column had rain or drizzle as a value, and 0 if not.

At last, two new columns categorizing the rain and wind intensity were created, as those prove to be important when evaluating the parking occupancy (Lijbers, 2016). The first column added was a column categorizing the rain information, resulting in a new column called *rain_intensity* that takes the values of the *weather_main* column and *weather_description* column into consideration. So, if the *weather_main* is equal to Drizzle, then this intensity is valued at 1, if the value is Rain, the *weather_description* column would be used to define the intensity of the Rain. If the *weather_description* is light, then the intensity is 2, normal intensity results in value 3, and heavy intensity is valued by 4. Finally, the *wind_intensity* column was created and, taking into consideration that the *wind_speed* values are not higher then 50.4 km/h, the *wind_intensity* column was created up to those values, while considering the Beaufort Scale, as it can be seen in Table 4.5. After that, the *wind_speed* column was removed.

TABLE 4.5: Wind speed categorization following the Beaufort Scale.

Wind speed (km/h)	Wind description	Wind intensity
≤ 2	Calm	0
≤ 5	Light air	1
≤ 11	Light breeze	2
≤ 19	Gentle breeze	3
≤ 28	Moderate breeze	4
≤ 38	Fresh breeze	5
≤ 49	Strong breeze	6
≤ 61	Moderate gale	7

To conclude, there are 3 different datasets with 2952 rows, one for each parking lot, where the Park 1 dataset has a total of 28 columns, the Park 2 a total of 26 columns and lastly, Park 3 dataset with 25 columns.

4.3 Data Evaluation

For this section, the essential focus is to carry out an extensive analysis of the previous data, in order to understand the variation in the parking lots occupancy rates and what type of features most influence those occupancy rates. In order to perform this analysis, tools like Microsoft Excel and Python language (using libraries such as Pandas and Numpy) were used, and for graphical presentation, Microsoft Excel was used again and Python's Matplotlib library.

To understand which features turn out to be more important and identify which of those have a bigger influence on the occupation rate inside the parking lots two techniques were used. The first is applied to binary features, where two values are compared graphically, as a way of concluding the impact that these values have and the second technique consists of analysing the correlation between the feature and the *occupation_rate*.

Correlation can give us a relationship between two values indicating that as one variable changes in value, the other variable tends to change its value in a specific direction. In this case, the Pearson's Correlation Coefficients were used, where the correlation coefficient value can range between -1 and 1, measuring both the strength and direction of the linear relationship between two continuous variables. Strength reveals that the greater the absolute the correlation coefficient is, the stronger the relationship is, meaning that as one value changes, the other will also change, where a coefficient of zero represents no linear relationship. As for direction, the sign of the correlation coefficient reveals the direction of the relationship, where positive coefficients indicate that when a value increases, the value of the other variable also tends to increase, and when the coefficient is negative it means that as one variable increases the other tends to decrease.

During this analysis, it is important to note that the occupation of the parks can reach, or even exceed, the occupation rate more than the times reflected, since the data obtained only allows us to know the occupation at the exact moment when the measure is taken, thus ignoring values that have occurred between measures.

It is also important to know that one important factor that may over-saturate the parking lot is the events on the surroundings of the location of the park (Xiao et al., 2018), so during this analysis there is a focus on looking for some outliers measurements and identify their cause. Also, if rich historical data can be implemented and information regarding the events is known in advance, the prediction can better adjust itself to take into account those special occasions, so it is necessary to understand when did those events occur in the historical data and identify them.

4.3.1 Over Time Occupation

In this section the occupancy from the rotation, covenants and total vehicles over time in the parking lots being studied are analyzed.

The over time occupation in Park 1 can be seen for the months of October, November, December and January on Appendix B. The appendix shows that the number of rotation vehicles remains relatively stable over the months having on average 200 vehicles during the workdays and less than 20 vehicles on weekends, except in special cases. In terms of covenants vehicles, the same pattern occurs for all months, where the workdays show an approximate total of 600 vehicles, around 500 more vehicles than in the weekends, that usually have around 100 vehicles. The maximum of covenants vehicles was reached on 14/01/2019 at 12:00 with a total of 706 and the maximum of rotation vehicles could be seen on 29/12/2018 at 18:00 with 686. This park is essentially composed of covenant type vehicles, representing on average 70%-80% of the occupation of the park. Finally, there was an increase of the total occupation from October to January, mostly due to the covenants vehicles which increased by 50, as the number of rotating vehicles remains very similar to the initial number.

The Appendix C shows us the occupation rate on the Park 2 over time for the covenants and rotation vehicles, and the combination of the two types of vehicle, during the months of October, November, December and January. The maximum rotation vehicles present in the parking lot at one point in time was 239 on 21/12/2018 at 22:00 and the maximum of covenants could be seen on two occasions, with a total of 185 vehicles on 05/12/2018 at 22:00 and on 27/11/2018 at 23:00. Unlike Park 1, Park 2 is mostly composed of vehicles of the type of rotation, being responsible for 65% of the park occupation most of the time. The total number of covenant vehicles has stayed relatively stable values over time, with an average value of 100 vehicles during the weekdays and around 20 during the weekends. For the rotation vehicles a pattern can be seen for all days of the week, where during the weekends there is an average of 30 vehicles and 200 for the rotation vehicles. Special cases occur that influence the occupation of the park, especially in the month of December.

Appendix D shows that there is a repetitive pattern in Park 3, where once again the days of the week present a higher affluence when compared to the weekends. On average, the days of the week present a maximum of 130 vehicles in the park, where around 70 are covenants and the rest of the rotation type. For the weekends, there is a lower affluence, having on average a maximum of 40 vehicles on Saturday, with approximately 10 of covenants and 30 of rotation. During the Sunday, a smaller occupation can be verified, with only 10 vehicles in the park. The maximum value of rotation vehicles in the park was 79, on December 11 at 15:00 and the maximum value of covenants vehicles was 81 also at 15:00 on December 17. Park 3 is the only park that presents the highest proportionality between the two types of vehicles, keeping most of the time a balance around 50% for each type of vehicle.

As it can be seen, the occupation in the parking lots reveals to have a very strong trend along the days of the week during the four months, where the most crowded days are the weekdays, generally reaching values around 80% and 90% of occupation, and sometimes very close to the total occupation or even reaching it. On weekends there is a big drop in the occupation rate, generally around a 10% of occupation for all parks, except for some cases that are explored later. So, it can be concluded that the occupancy rates along the parks remain very similar.

It can also be concluded, in a general way, that holidays, mainly those of importance 2, as well as vacation period have a strong influence on the occupation of the parks, but those will be analyzed further on. There is also a clear difference in the occupation of the parks between weekdays and weekends which is essential to highlight, as well as some difference between the days and months. These scenarios are analysed in detail in the following section for each of the parks, in order to understand if the impacts are greater in the vehicles of type rotation, covenants or in the total occupation.

4.3.2 Outliers Data

Outliers days could happen due to various causes and in this section the focus is to identify the days when these cases occurred and finding out why some of these cases happened, by framing them into two possible cases, invalid measurements or events.

4.3.2.1 Events for Park 1

For the month of October, it can be seen small peaks on the rotation vehicles for the days of 18/10/2018 and from 23/10/2018 to 25/10/2018, with around an extra 100 vehicles, probably due to the ”Doc Lisboa” event, being this a documentary film festival in Lisbon, happening during October 18, 2018 to October 28, 2018. It can also be seen an increase of about 100 rotation vehicles on the second weekend of the month (13 and 14 October) when comparing with the other weekends in the month.

In the case of the month of November, it was identified on the 16th an increase in the occupation of Park 1 reaching a maximum of 285 rotation vehicles, an increase of 58% when comparing to a normal Friday in the month of November. That could be due to the ”LEFFEST” event (Lisbon & Sintra Film Festival) a film festival including areas such as literature, music, visual arts, among others, which took place during the days 16 of November to 25 of November in various locations of Lisbon and Sintra.

During the month of December, it can be seen this increase specifically on the weekends before Christmas, where the number of rotation vehicles increase around 600%-1300%. This increase can be justified by the ”Wonderland Lisboa” event, right beside Park 1 and also, the first two Saturdays are holidays of importance 2, that can consequently increase the allocation to the event and can thus contribute to the increase in the number of rotation vehicles. Between 24 and 25 of December there is a decrease, having only around 150 covenant vehicles, that could be due to the Christmas holiday. The following week it can also be seen a small decrease of about 100-200 covenants vehicles when compared to similar periods of other months. For the last weekend of the month, a high value of rotation vehicles can be seen, reaching a maximum of 686 vehicles at 12:00 of 29 December, this maximum could be due to the ”Corrida São Silvestre de Lisboa 2018” event. This event is a race in which the course of the race crosses the centre of the Portuguese capital, with a starting point on Avenida da Liberdade, one of the roads previously selected as influential for Park 1.

Lastly, on the month of January a low occupation on the first day can be seen, with a maximum of 416 vehicles, much lower when comparing to the average maximum of 800-900 vehicles during a weekday for the month of January, but justified by the fact that this is the holiday "Dia de Ano-Novo", a holiday of importance 2. However, the total number of rotation vehicles was higher than normal, reaching almost 300 vehicles, this can be justified by the fact that on January 1st there are several events to celebrate the new year. On the 29th of January a maximum of 316 rotation vehicles can be verified, an increase of 27% when comparing to a normal Tuesday of January, and on the 30th of January there is a maximum of 347 rotation vehicles, revealing an increase of 44% of the rotation vehicles, that may be due to the ”Building the Future: Ativar Portugal” event that happened in the Pavilhão Carlos Lopes.

4.3.2.2 Invalid Measurements for Park 1

For the Park 1 several invalid measurements were found, the first was a large peak during day 7 (Sunday) October at 22:00, of about 315 covenants vehicles, more 202 vehicles then the previous measurement and more 209 than the measurement after, probably representing an erroneous measure.

In the month of January some strange situations were also noticed, namely drastic falls and climbs within an hour in the total number of covenants vehicles from one measure to the next, namely on 2/01 at 21:00, on 5/01 at 17:00 and at 20:00, on 12/01 at 20:00, on 16/01 at 3:00, 4:00 and 7:00, on 28/01 at 13:00 and finally on 30/01 at 20:00. Park 1 data also showed one measurement in 29 January, 2019 at 16:00 where the occupation rate was 119%, with 319 rotation and 969 covenants vehicles in the parking lot.

Another strange situation was identified on Monday of 19th November, shows occupation levels much lower than a normal Monday with a maximum of 28 vehicles and on 17th November (Saturday) an unexpected quantity of vehicles in the parking lot can be seen, reaching more than 800 vehicles, implying that there was an exchange of results between the two days, so they were switched.

4.3.2.3 Events for Park 2

In October only one occasion was identified, more specifically on the 13th of October, there was an increased value of the total occupation when compared with other days, probably by some event in the surroundings.

The month of December, once again shows a big difference when compared to the occupation pattern of the other months, especially during the weekends for the rotation vehicles. The month of December starts with two holidays that showed big impacts on the occupation of the parking lot. On the first holiday (01/12/2018), there is a big allocation of rotation vehicles for a Saturday with an increase of 650% for the rotation vehicles. On the second day the same effect can be seen, having 183 rotation vehicles at a certain point from the 206 present in the park, resulting on a increase of around 800% for the type rotation. The following day (02/12/2018), on 9 December (Sunday) and on the 15th show high values of occupation thanks once again to the rotation vehicles. Those values of occupation could also be due to the "Wonderland Lisboa" event, still very close to the Park 2 or the event "Natal em Lisboa" where a bunch of Christmas concerts happen around Lisbon between 1/12 to 23/12. In the last week of December the impact of Christmas for the total occupation of the parking lot can be verified, where there was no more than 31 vehicles at a time during 24th and 25th. From the 26th to 28th an increase of the parking occupation can be seen, but still very low when comparing to the normal cases, probably due to the fact that a lot of people take a vacation during this period. At 29th a large peek of rotation vehicles for a Saturday can be checked, probably due to the "Corrida São Silvestre de Lisboa" 2018.

4.3.2.4 Invalid Measurements for Park 2

For Park 2 one invalid measurement was found with 618% occupation (113 rotation vehicles and 1963 covenants) on the 16 October, 2018 at 16:00. Some strange measurements on 20 November at 7:00 and on the 27th at 23:00 were also seen, where sudden changes in total occupancy values can be found, thus showing that those measures have the potential to be wrong. Three instances where the occupation rate values were suspicious were verified, namely on the 21 December at 22:00 and on 31 December at 15:00 and at 16:00, where once again sudden changes in total occupancy values can be seen. Another invalid measurement was found where there was a 101% occupation rate, having a total of 337 vehicles inside on the 25 January, 2019 at 12:00.

4.3.2.5 Events for Park 3

In December there is a big impact on Sunday the 9th, with an increase of around 100% when comparing with other Sundays, probably due to one of the events in

December, namely the ”Wonderland Lisboa” or ”Natal em Lisboa”. The impact of Christmas on the park occupancy can also be seen, where the 24th of December had low occupation values for a Monday, no more than 34 vehicles and where the 25th of December had most of the time 10 vehicles in the parking lot, much lower when compared to other Tuesday values. The following days the same exact values can be seen for the same date and time measurements, which led us to conclude that these measurements are repeated and in turn erroneous. That said, the only measure that was counted as valid was the first, on the 26th, which shows a maximum occupancy value of 91 vehicles, this being lower when compared to other Wednesdays. Although there is no information on the last days of December, it could be concluded, by comparing with the other parks, that the allocation to this park also decreases during this period, and this may be due to the festive period of this month which may lead to holidays by users.

4.3.2.6 Invalid Measurements for Park 3

An invalid measurement was found where the occupation rate was 399%, with a total of 57 covenants and 557 rotation, on 31 October, 2018 at 14:00. Another two wrong measurements were seen, one with 477% of parking occupancy, with 69 covenants and 665 from rotation, on 5 November, 2018 at 16:00 and the second with 164%, with a total of 241 covenants and 11 rotation, on 27 November, 2018 at 20:00. Another occasion can be seen on day 11 November of 2018 at 1:00 there are no values for covenants too and lastly, on 12 November, 2018 at 0:00 there are no measurements for rotation and covenants. Some strange behaviours of probably wrong measures can be observed due to sudden shifts of the occupancy values, namely on the 9th December at 12:00, on the 17th December at 8:00 and on the 21st December at 9:00. On the last 5 days of the month of December (27, 28, 29, 30 and 31) the data was missing for the total occupation on the parking lot. Some minor inaccuracies on Park 3 can be checked, namely on day 10 January, 2019 where there are no measurements for the covenants column.

Taking into account the events that have been identified and that influence each of the parking lot, a column called *flag_ event* was added that has a value of 1 during the dates of the events that have been identified and a value of 0 for the remaining days.

4.3.3 Occupation rates throughout the week

As seen in the previous analyses, the occupancy of each parking lot differs greatly from a weekday to the weekend, so in this section the average occupancy rate of the parking lots was analyzed over the various days of the week, so that the magnitude of the influence that a weekend has on the occupancy rate can be concluded.

4.3.3.1 Park 1

In Figure 4.2 the average occupation rate throughout the weekdays on Park 1 can be seen. By analyzing this figure, it can be seen that there is a clear difference in the occupation rates between the normal weekdays and the weekends, where the weekdays reach values higher than the 60% and weekends generally not higher than 25%. Wednesday and Thursday prove to be the busiest business days and, in contrast, Monday the least. The weekends occupation rate have a different pattern than the normal weekdays, since the maximum occupation rate is generally reached after 17:00 on weekends and in the weekdays the maximum is reached between 10:00 to 16:00.

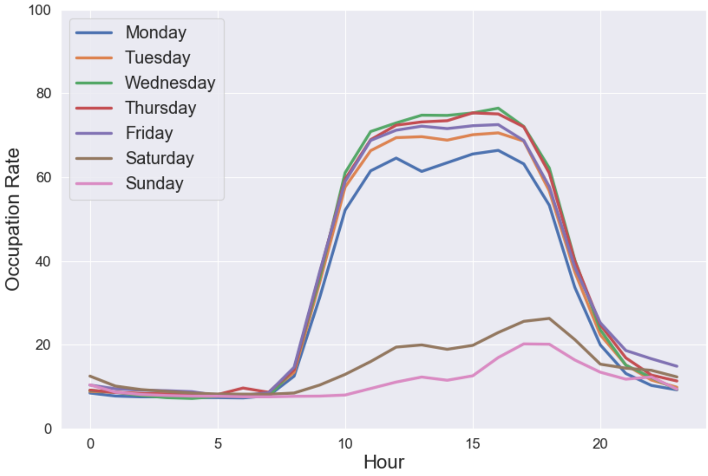


Figure 4.2: Average occupation rate for Park 1 on the various weekdays.

4.3.3.2 Park 2

Seeing Figure 4.3 there is a clear difference between the weekdays and the weekends on Park 2, where the first ones reach values around the 80%, and the later ones reach no more than 30%. There is a big difference between the weekends days, with Saturday reaching occupation levels around the 30% and Sunday reaching a maximum of 15%. During the weekdays, there are very similar values for the occupation rate, with Wednesday and Thursday being the busiest days.

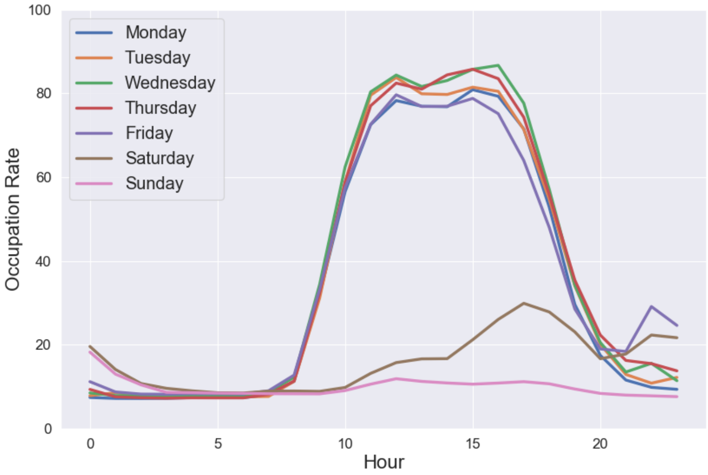


Figure 4.3: Average occupation rate for Park 2 on the various weekdays.

4.3.3.3 Park 3

Lastly, at Figure 4.4 similar patterns can be checked to those in Park 2 on Park 3. A big difference can be seen between the weekdays and the weekends, mainly between Wednesday/Thursday and Sunday, where first two reach values of occupancy higher than 80% and Sunday reaching occupation values of only 20%. The average occupation of Saturdays also shows a big difference to the occupation values in Sunday, with a maximum value of 30% for Saturday and 10% for Sunday.

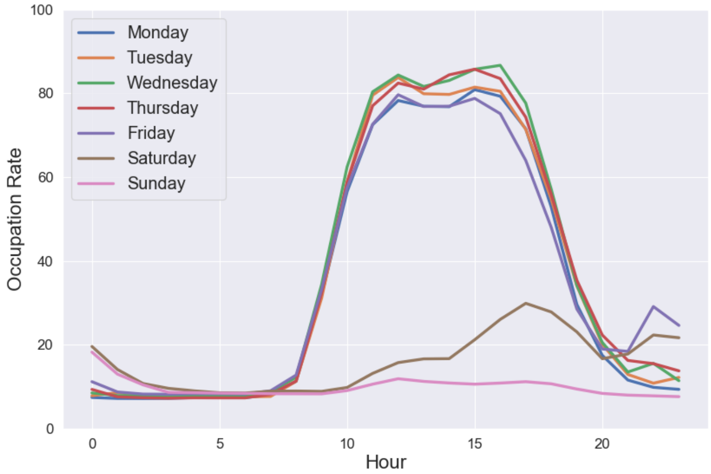


FIGURE 4.4: Average occupation rate for Park 3 on the various weekdays.

4.3.4 Holidays

Once again the previous analyses makes it possible to verify that holidays have an enormous impact on the occupation of parks. In Figures 4.5, 4.6 and 4.7, the average occupation of the parks being studied for a no public holiday day, a holiday of importance 1 and a holiday of importance 2, respectively, can be visualized.

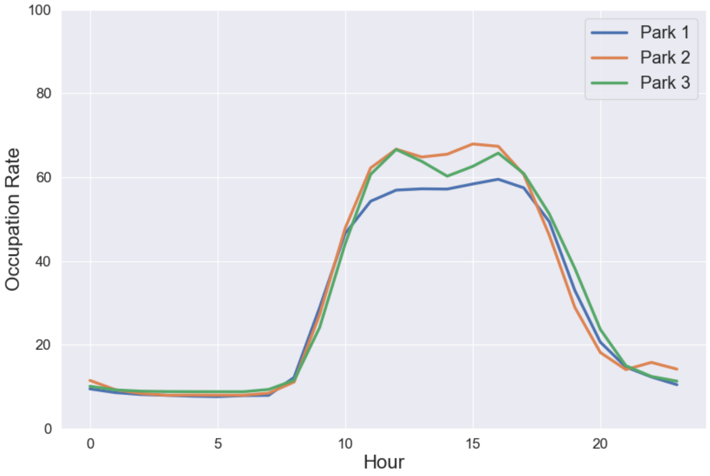


FIGURE 4.5: Average occupation rate on a Holiday of value 0 importance.

For a normal day (in here all of the weekdays and the weekends with value 0 on the *flag_holiday* were counted), seen in the Figure 4.5, the maximum average occupancy rate is around the 60 percent mark. The occupation levels starts to increase for the three parking lots at 8:00 and keeps increasing until the 12:00, where Park 1 reaches a maximum of 57% and Park 2 and 3 a maximum of 67%. Park 2 also reaches this maximum at the 15:00. The occupancy levels are kept high until 16:00 when it starts to decrease, reaching the 10% occupancy levels at 21:00.

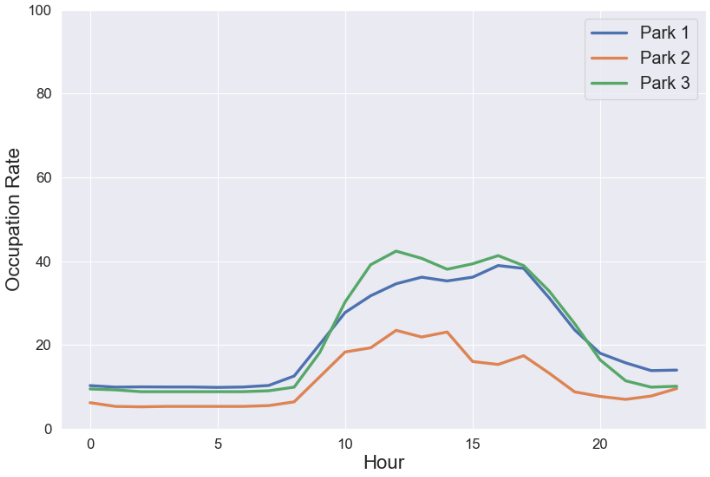


FIGURE 4.6: Average occupation rate on a Holiday of value 1 importance.

In the Figure 4.6, the average occupation rate for the Holiday with an importance value 1 can be seen. The occupation rate keeps a value of around 10% from the 0:00 until the 8:00 where it starts to increase until the 13:00. For Park 1 this increase keeps on happening until the 17:00, reaching the maximum occupation of 39%, where after that it decreases to 16%. In the case of Park 2 the lowest occupancy can be seen, reaching a maximum of 22%, maintaining values around the maximum until the 17:00, where it starts to decrease to 7% at 20:00. Finally, Park 3 is the least affected by this type of holiday, where the occupancy reaches values of 43%, keeping these figures until 17:00, decreasing to 10% by 21:00.

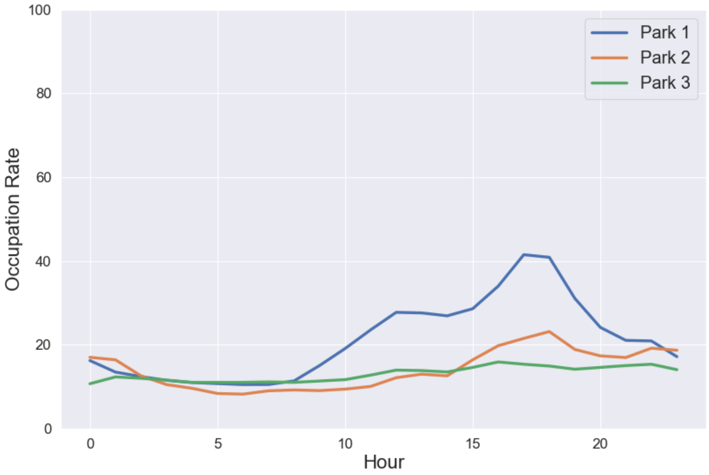


Figure 4.7: Average occupation rate on a Holiday of value 2 importance.

The occupancy in the parking lots during a Holiday of importance 2 can be seen in the Figure 4.6 and reveals to have a big impact, specifically on Park 3. The parking lots occupation rate starts at around the 10%, where for Park 1 occupation levels start to increase around the 8:00 and reaching the maximum 42% around 17:00. Park 2 occupancy only starts to increase at 15:00 until the 18:00 with a maximum of 23%. Park 3 is the most affected by the holidays with importance level 2, maintaining around the 14% occupation rate all day long.

The holidays have a big impact on all parking lots occupation, having a difference in the maximum parking occupancy of about 20% on the holidays type 1 and 15% on the holidays type 2 when comparing with a normal day for Park 1, being this the least affected park out of the three by the holidays. In the case of Park 2, when comparing with a not public holiday the difference can be up to 43% for type 1 and 44% for type 2, proving to have a bigger impact than onto Park 1. For Park 3 the maximum occupancy reached on a normal day was 67%, having a difference of 24% for type 1 and 53% for the type 2 holiday, meaning that the type 1 does not have that much impact, but the type 2 does have a huge impact.

4.3.5 Vacation Period

As previously stated, a vacation period was defined between the dates 22/12/2018 and 02/01/2019. To help analyse the importance of this period and the differences to a normal period the Figure 4.8 and Figure 4.9 are available.

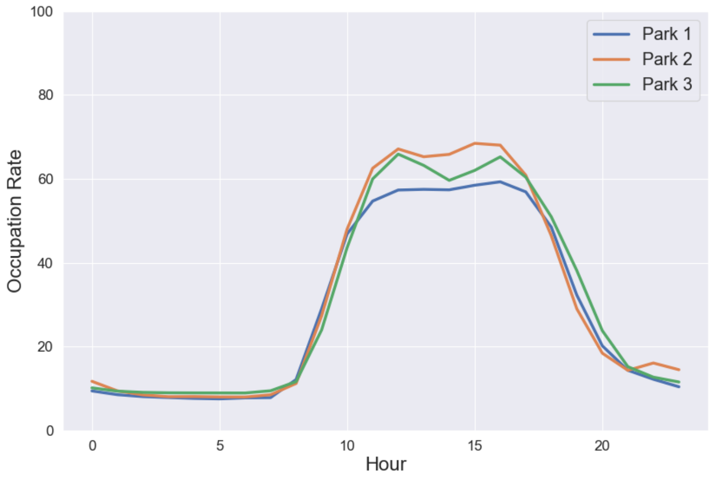


FIGURE 4.8: Average occupation rate during the non vacation period.

From the analysis of Figure 4.8, the average occupation rate for all three parking lots where the vacation period flag is equal to 0 can be visualized. The occupation starts to increase around 7:00 am. The occupancy of Park 1 and 2 stays relatively the same until 17:00 and for Park 3 there are two peaks, one at 12:00 and other at 16:00. The Park 1 reaches an average maximum occupation value of 58%, Park 2 reaches 68% and Park 3 66%. From 17:00 the park occupancy of all three parking lots decreases to reach values around 15% park occupation.

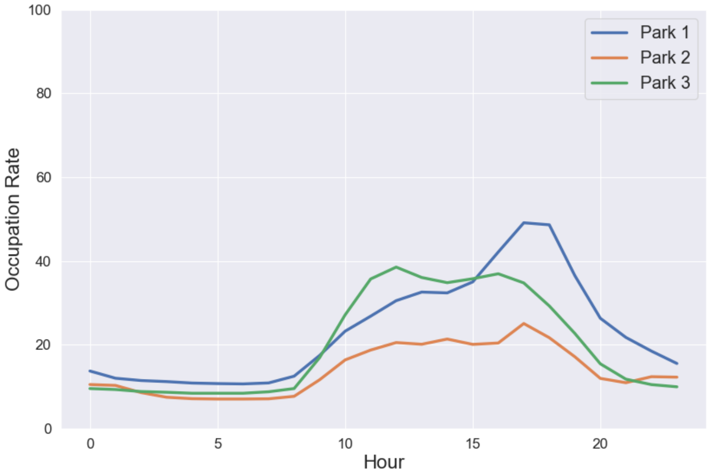


FIGURE 4.9: Average occupation rate during the vacation period.

The Figure 4.9 shows the average occupancy of the parking lots during the vacation period. In this case the occupation starts to increase at 8:00 am, one hour later than during a normal period, and increases up to 48% for Park 1 at 18:00, up to 25% for Park 2 at 17:00 and up to 39% at 12:00 for Park 3. Park 1 maintains the highest occupancy value for over an hour and then it starts to decrease until reaching the 15% occupation rate at 23:00. For Park 2 has the smallest occupation rate, reaching only a 25% occupation rate and decreasing immediately after that, until reaching the 10% occupation mark at 20:00. Finally, Park 3 shows two high points of occupation, one at 12:00 reaching 39% and at 16:00 reaching 38%.

For the three parking lots the occupation rate is affected by this vacation period, especially Park 2. The least affected park is Park 1, having only a 10% occupation difference for the maximum values of occupation when comparing with a normal day. Park 2 shows a difference of 43% between the highest occupation values in a normal period and a vacation period. Lastly, Park 3 showed a difference of 27% between the highest points of occupation on each period, revealing to be the second most impacted parking lot by the vacation period.

4.3.6 Park Occupancy and Weather Data

In this section, the impact that weather conditions have on each of the three parking lots is analyzed, namely the temperature, humidity, wind and rain intensity with correlation values, and also the impact of rain and fog graphically.

4.3.6.1 Temperature Correlation with Occupation Rate

Here the impact temperature information has in the occupancy of each parking lot is evaluated, namely the current temperature (*temp*), seen by Figure 4.10 temperature and in Table 4.6 the respective correlation results for the temperature feature.

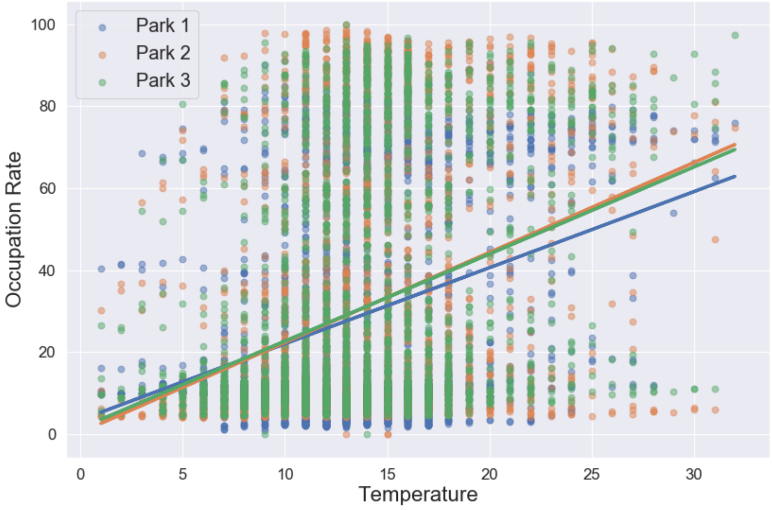


FIGURE 4.10: Correlation between the temperature feature and the occupation rate from the parks.

The Figure 4.10 shows that there is a weak positive linear correlation between the temperature features and the occupation rate, showing very scattered data and outliers, since the increase of a temperature variable does not mean the increase of the occupation rate, or the inverse.

TABLE 4.6: Correlation results of the temp feature with the occupation rate.

Park	temp
Park 1	0.31
Park 2	0.32
Park 3	0.33

The correlation values using the Pearson's Correlation can be seen in Table 4.6 and are very low, thus proving the conclusions drawn when analyzing the data of the previous figure. The *temp* feature has a 0.31 correlation value for Park 1, 0.32 for Park 2 and 0.33 in the case of Park 3.

4.3.6.2 Humidity Correlation with Occupation Rate

In Figure 4.11 the impact of the humidity feature (*humidity*) in the occupancy of the three parking lots can be seen and Table 4.11 has the Pearson's Correlation Coefficients for the humidity feature.

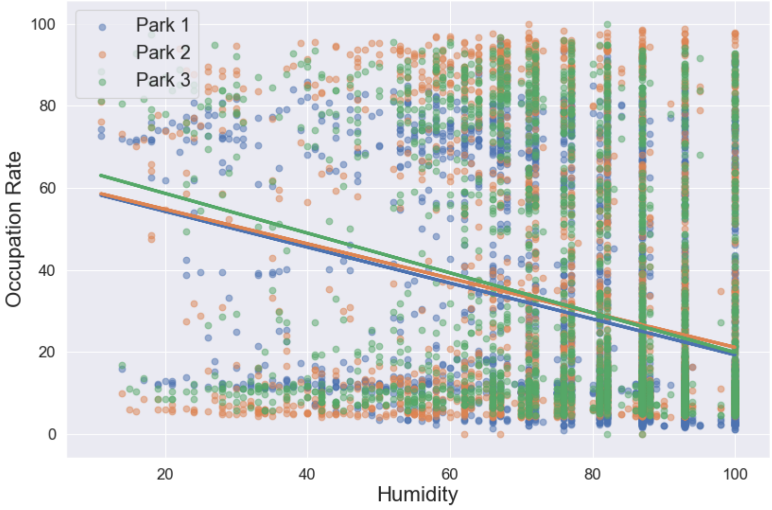


FIGURE 4.11: Correlation between the humidity feature and the occupation rate from the parks.

Figure 4.11 shows that there is a weak negative linear correlation between the humidity feature and the occupation rate, meaning that the increase of the humidity variable, generally means the decrease of the occupation rate and that the decrease of humidity value results in the increase of the occupation rate.

Table 4.7: Correlation results of the humidity feature with the occupation rate.

Park	humidity
Park 1	-0.28
Park 2	-0.24
Park 3	-0.27

Table 4.7 shows that the values are weak in terms of strength, not exceeding the -0.28 value, representing the value for Park 1. Park 3 obtained the second highest value with -0.27 and lastly, Park 2 with only -0.24 for the correlation coefficient value, concluding that the humidity is not an influential feature to the parking occupancy on the three parking lots.

4.3.6.3 Wind Intensity Correlation with Occupation Rate

Here the impact of the of the wind intensity column (*wind_intensity*) has in the occupancy of each parking lot is analyzed, seen by the Figure 4.12 and by the Table 4.8.

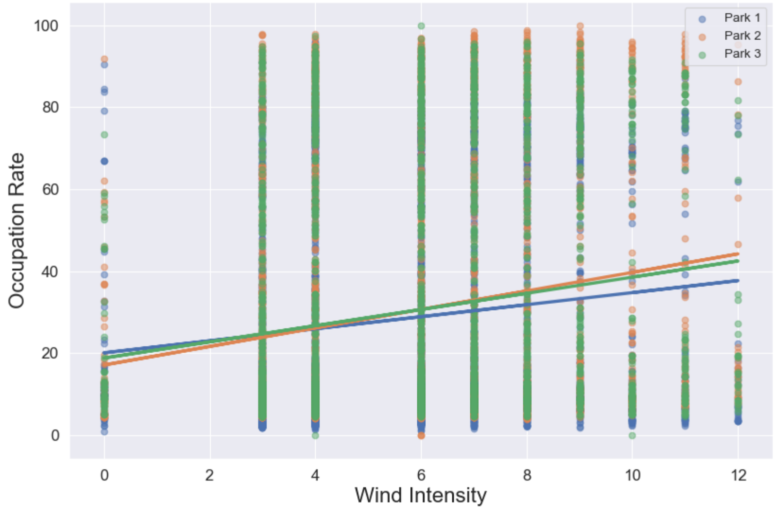


FIGURE 4.12: Correlation between the wind intensity feature and the occupation rate from the parks.

A weak positive linear correlation for all parking lots between the wind intensity feature and the occupation rate can be seen by analyzing the Figure 4.12, concluding that the wind intensity does not influence the occupation of parking lots.

TABLE 4.8: Correlation results of the wind_intensity feature with the occupation rate.

Park	wind_intensity
Park 1	0.13
Park 2	0.18
Park 3	0.16

The results obtained in Table 4.8 shows that the wind intensity does not show high values of correlation. Park 2 had the biggest coefficient value with 0.18, and Park 3 had the second highest value with 0.16 and finally, Park 1 with 0.13. The values are low, thus reinforcing the previous conclusion, meaning that there is no real correlation between both variables.

4.3.6.4 Rain Intensity Correlation with Occupation Rate

In Figure 4.13 the impact of the rain intensity feature (*rain_intensity*) has in the occupancy of the parking lots can be seen and in Table 4.9 the Pearson's Correlation Coefficients between the rain intensity feature and the occupation rate are presented.

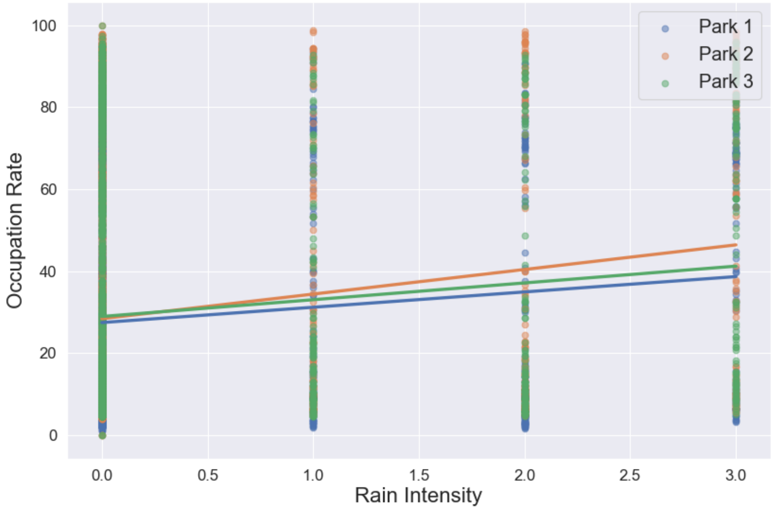


FIGURE 4.13: Correlation between the rain intensity feature and the occupation rate from the parks.

The Figure 4.13 shows that the correlation between the rain intensity and the occupation is a positive linear correlation, but with very low strength.

TABLE 4.9: Correlation results of the rain_intensity feature with the occupation rate.

Park	rain_intensity
Park 1	0.09
Park 2	0.14
Park 3	0.12

Checking the results obtained in Table 4.9, shows that the values for the Correlation Coefficient are very low, being in agreement with the previous conclusions drawn in Figure 4.13. Taking this into consideration, the feature *rain_intensity* is not influential on the occupation rate for all three parking lots.

4.3.6.5 Rain versus No rain

In Figure 4.14 and Figure 4.15, the average occupation of the parks being studied by comparing a no rain day and a rain day, where the *flag_rain* is equal to 1, can be seen. With the *flag_rain* valued as 1 there are a total of 361 measurements.

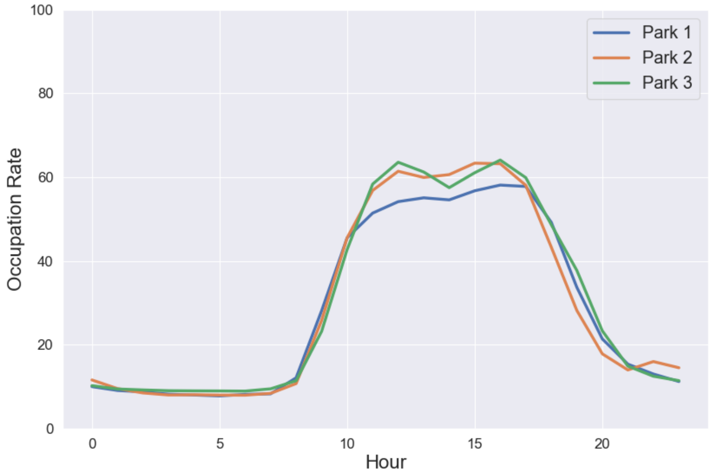


Figure 4.14: Average occupation rate for a day without rain.

The Figure 4.14 shows the average occupation rate over time for a day with no rain and showing that the occupation rate maintains values pretty close to those of Figure 4.5. Park 1 occupancy keeps increasing until it reaches a maximum total occupation of 58% at 15:00, after that the occupation levels start to decrease until reaching around the 10% occupancy mark. For Park 2 the same pattern occurs, where the occupation rate increases up to 63% at the 15:00 and starts to decrease from that point on, reaching a total of 14% occupation rate at the end of the day. Lastly, Park 3 starts the day with a 9% occupation rate, starting to increase until the 12:00 with a total of 62% occupation rate, decreasing a little for the next three hours and then reaching the maximum occupation of 63% at 16:00. After that the occupation keeps on decreasing until it reaches the total of 10% occupation level at 21:00.

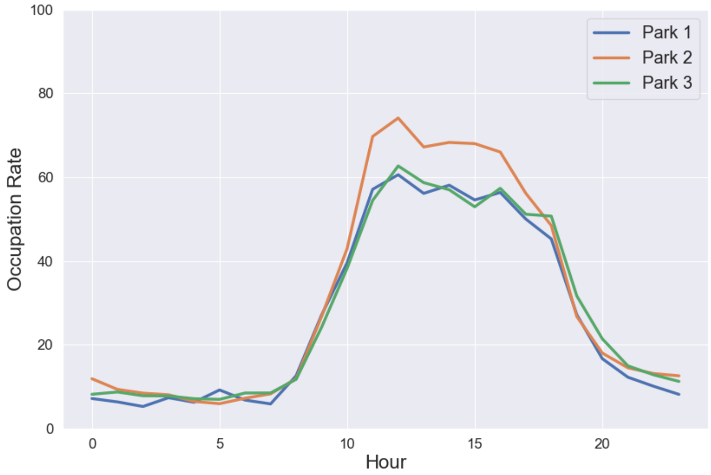


FIGURE 4.15: Average occupation rate for a day with rain.

In Figure 4.15 the average occupation rate on a rainy day for the three parking lots can be verified. Analyzing this figure shows that the maximum occupations are reached earlier than in a day without rain, where Park 2 was the one with a higher occupancy reaching 74% of occupancy at 12:00, with Park 1 and 3 having very similar occupation patterns and reaching the maximum also at 12:00 with Park 2 having 62% and with Park 1 having 60%. After that, both parks maintained values around the 60% occupation mark until 16:00 where the occupation for all of the parks started to decrease until reaching around the 10% occupation mark at 22:00.

So, in conclusion, a non-rainy day has a different occupation trend then a rainy day, so it is important to take this information into consideration, because all three parks are influenced by this feature.

4.3.6.6 Fog versus No fog

In the Figure 4.16 and Figure 4.17 the influence of the fog on the average occupation of the parks for a day with no fog and day with fog, meaning that the column *flag_fog* is equal to 1, can be visualized. There are 336 measurements for each of the parking lots datasets where the *flag_fog* is equal to 1.

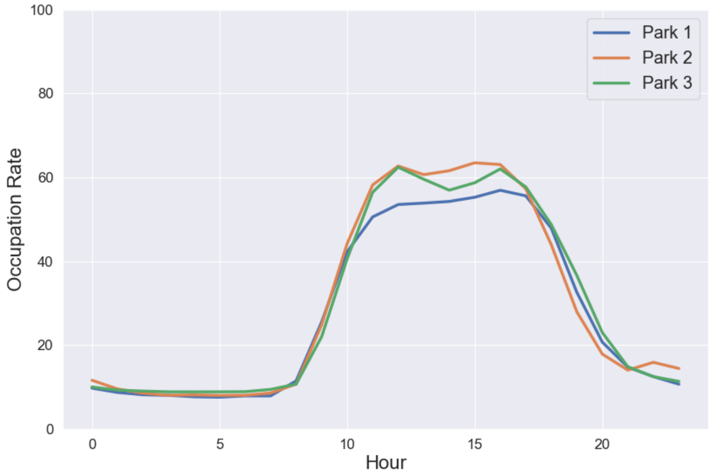


FIGURE 4.16: Average occupation rate for a day without fog.

Figure 4.16 shows that the occupation rate for a day with no fog is quite similar to the occupation on the Figure 4.14. There is a continuous occupation rate for all three parking lots until the 8:00 where it starts to increase until the 12:00 where Park 1 reaches 54% occupation and Park 2 and 3 reaches 62%. After this the occupancy levels stays relatively stable for all three parks, until 16:00 where it starts to decrease until it reaches an occupancy rate of around 10 percent by 22:00.

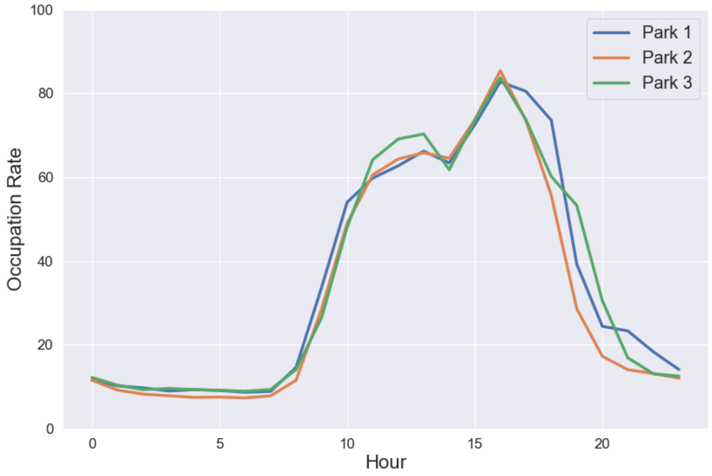


FIGURE 4.17: Average occupation rate for a day with fog.

On the Figure 4.17 the impact that the fog has on the parks can be verified. Right at first glance, the occupation rate reaches higher values than on a no fog day. Initially the occupancy levels increase from the 8:00 to the 13:00, reaching values of 66% for Park 1 and Park 2, and 70% for Park 3. On the next measurement, there is a small decrease on the occupancy for all of the parking lots, increasing right after until reaching the maximum occupancy for each park. At 16:00, the maximum occupancy for each parking lot is reached, where Park 2 had the highest occupation rate with 85%, Park 3 with 84% and lastly Park 1 with 83%. After, the parking occupancy starts to decrease until it reaches the 10% occupancy mark for all three parks at 23:00.

With this, the occupation pattern for the three parking lots is quite different on a day with fog when comparing to a day with no fog, concluding that this information is influential for the state of each park, as the parking occupancy increases.

Taking into account all the previous conclusions, not every weather information in the surroundings of the parking lots have an influence in the occupancy, since these parking lots are categorized as office parking lots, this means that regardless of the weather conditions, people need to move to their place of work, concluding that the influence of the weather is not very strong on the occupancy rate of each park. So, taking into account the previous analyses, the weather variables with the greatest influence on the final result are the *flag_rain* and *flag_fog*.

4.3.7 Park Occupancy and Traffic Data

In this section the impact of traffic conditions have on each of the three parking lots are analysed, as well as the correlation values between the occupation rate and the average time it takes to travel through the road, obtained by using the previously identified correlation method of Pearson.

4.3.7.1 Park 1

In the Appendix E the correlation of the average wait time in a road with the occupation of Park 1 can be verified. The correlation coefficient values obtained with the Pearson's Correlation method in the Table 4.10 can also be checked.

TABLE 4.10: Correlation results of the average_time features for the surrounding roads of the Park 1 with the occupation rate.

Road	average_time
Praça Marquês de Pombal	0.69
Túnel do Marquês de Pombal	0.27
Rua Braamcamp	0.62
Avenida da Liberdade	0.60
Avenida Duque de Loulé	0.66
Rua Joaquim António de Aguiar	0.60
Avenida António Augusto de Aguiar	0.44
Alameda Edgar Cardoso	0.46
Avenida Fontes Pereira de Melo	0.56

Appendix E shows that some of the roads have better correlation with the occupation rate of the parking lots. Checking all of the road segments, there is a positive linear correlation where some of them show a stronger strength, but the values seem to be very scattered, as well as varied outliers occurrences. However, it is possible to identify by checking Table 4.10 that the roads with a higher strength of correlation seen are the Praça Marquês de Pombal, Avenida Duque de Loulé, Rua Braacamp, Avenida da Liberdade, Rua Joaquim António de Aguiar and Avenida Fontes Pereira de Melo, with values higher then 0.56, showing a large strength of association. These being the only streets taken into consideration from now on for this park.

4.3.7.2 Park 2

The scatter plot for the correlation of Park 2 occupation rate with the average wait time in the road in the surroundings of the parking lot can be seen in the Appendix F and the respective correlation values for each road segment can be checked in Table 4.11.

Table 4.11: Correlation results of the average_time features for the surrounding roads of the Park 2 with the occupation rate.

Road	average_time
Praça Marquês de Pombal	0.69
Rua Alexandre Herculano	0.51
Rua Castilho	0.67
Rua Braamcamp	0.57
Rua Mouzinho da Silveira	0.36
Rua Duque de Palmela	0.26
Túnel do Marquês de Pombal	0.25

Appendix F and Table 4.11 shows that the roads with higher correlation with the occupation of Park 2 are: Praça Marquês de Pombal, Rua Castilho, Rua Braamcamp and Rua Alexandre Herculano, as those are the roads showing values higher than 0.50, often declared as the minimum acceptable for a feature to be really influential, in this case the occupancy rate of the park, being once again the roads used from now on, when it comes to this park.

4.3.7.3 Park 3

In the Appendix G the influence of the average time to travel through the roads closer to the park with the occupation rate of Park 3 can be visualized, as well as the values of the coefficient correlation values obtained with the application of the Pearson's Correlation method in Table 4.12.

TABLE 4.12: Correlation results of the average_time features for the surrounding roads of the Park 3 with the occupation rate.

Road	average_time
Praça Marquês de Pombal	0.68
Avenida Engenheiro Duarte Pacheco	0.37
Rua Castilho	0.69
Rua Braamcamp	0.62
Rua Joaquim António de Aguiar	0.59
Túnel do Marquês de Pombal	0.28

By analyzing together the Appendix G and the Table 4.12 shows that the roads with greater influence on Park 3 are the Rua Castilho, Praça Marquês de Pombal, Rua Braamcamp and Rua Joaquim António de Aguiar, in order of highest correlation value. These were the streets with the highest correlation value, with Rua Joaquim António de Aguiar, the last to be selected, having a correlation value above 0.59 and Rua Castilho with 0.69 being the street with the highest value.

Having said that, by increasing the average length of time to cross some streets, the occupancy rate of the park increases as well, this means that traffic around the parking lot has a high correlation with the number of vehicles inside the park. The higher the movement outside the park, the higher the number of vehicles in the parking lot.

4.4 Data Treatment

In order to deal with data with wrong/missing measurements the same methods used in (Stolfi et al., 2019) were used, instead of simply removing those measurements, since the current amount of data is small. The first method focus on dealing with wrong daily measurements (wm_w), being those filled in with the average of the four days of the week preceding the wrong one when an entire day is wrong, that is, if the day 31st December (Monday) is wrong/missing, the average of the values of the four Mondays before (December 3rd, 10th, 17th and 24th) is applied, seen in Equation 4.1.

$$wm_w = \frac{wm_{w-1} + wm_{w-2} + wm_{w-3} + wm_{w-4}}{4}, w \in Weekdays \tag{4.1}$$

In the case that only one measure is wrong (wm_h), the average between the previous and posterior measurement is applied 4.2. This treatment has been applied for all the cases previously identified where there were sudden shifts of occupation values on the parking lot or the measurement was missing. For example, if on 31st December at 10:00 the measurement was identified as wrong or missing, it would be replace by the mean value between the measurement at 9:00 and the measurement at 11:00.

$$wm_h = \frac{wm_{h-1} + wm_{h+1}}{2}, h \in Hours \tag{4.2}$$

With those treatment techniques there was no need of removing any invalid measurement previously mentioned, ending with a total of 2952 rows for each dataset.

4.5 Data Selection

To conclude, there are some possible reasons for some of the features analyzed not having much more correlation with the park occupancy. Three reasons could justify this behaviour, namely the parking lot categorised as an office category park, where people have to move to their work and park the vehicle there often regardless of the weather, traffic and events. The second reason may be because the park is underground, which can cause the weather condition not to be so critical to the affluence of the park. In contrast, data from events in the vicinity of the park show to be quite useful when forecasting the occupation of the park, this is due to the fact that the occupation, mainly of the total number of rotation vehicles, changes considerably at the times when an event occurs in the vicinity. And the last reason is that the user with a covenant parks there regardless of the weather, traffic and events on the surroundings, since they have already paid a sum to secure a place in that park.

The data concluded to have been influential to the affluence of the parking lots categorized as an office park with covenant options can be seen in the past subsections. Features like events and holidays have big impact on the total occupation, just like the closest roads to the parking lot, meaning that those conclusions can be propagated to other parking lots with similar characteristics and same categorization.

So, taking into account all previous conclusions, the data to produce the final dataset and kept for the rest of the study is the following: *year*, *month*, *day*, *hour*, *flag_ holiday*, *weather_ description*, *weather_ main*, *flag_ event*, *flag_ vacationperiod*, *flag_ rain*, *flag_fog*, *dayofweek*, *occupation_ rate*, *flag_ weekend*, as well as the columns for each road selected. The covenants, rotation and total occupation columns were removed, as those columns were highly dependent and highly correlated with the value *occupation_ rate* used to predict and by leaving those columns, the training of the predictive models could be influenced.

Chapter 5

Predictive Model Development

This chapter consists of the development and testing of various predictive models, where initially a brief introduction to the algorithms used and the environment in which they were created is made. Tests of each predictive model are performed and in turn the results obtained are addressed. Finally, the predictive model that presented the best results is optimized, after which it is imported into the mobile application.

5.1 Chosen Algorithms

To build the predictive models, with the aim of predicting the total occupation of the parking lot, three algorithms have been chosen, specifically the Gradient Boosting Machine (GBM), Distributed Random Forest (DRF) and NN from the python library H2O (H2O.ai, 2019), which are usually used to deal with this type of problem. In the following steps a small overview of the algorithms used in this study are given considering the H2O documentation.

5.1.1 Gradient Boosting Machine (GBM)

Gradient Boosting Machine, that works either for a Regression and Classification, is a forward learning ensemble method, that works by working on an increasingly refined approximations to build better good predictive results, being a good option as those gave the best results in (Ionita et al., 2018). The implementation of GBM by H2O sequentially builds regression trees on all the features of the dataset in a fully distributed way, where each tree is build parallel, being beneficial to the performance of the model.

5.1.2 Distributed Random Forest (DRF)

Distributed Random Forest, is a powerful classification and regression tool, that generates a forest of classification or regression trees when data is given, rather than a single classification or regression tree. Each of the trees is a weak learner built on a subset of rows and columns and the more trees there are, the smaller the variance. For the two cases of regression and classification, this technique takes the average prediction over all of the trees to make a final prediction, whether predicting for a class or numeric value. In (Zheng et al., 2015) the Regression Trees had better results and less computationally needs comparing to SVR and NN.

5.1.3 Neural Networks (NN)

The H2O's Neural Networks are based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation. A feedforward artificial neural network model, also known as deep neural network or multi-layer perceptron, is the most common type of Deep Neural Network and is the type of neural network implemented here. The NN is a good solution for a time series prediction like our problem of parking occupancy, as concluded in (E. I. Vlahogianni et al., 2016),

5.2 Prediction Models Preparation

The previously identified algorithms were applied to each parking lot dataset with the features concluded in the previous chapter. Those datasets where divided in 70% to train data, 15% to test data and 15% to validation data using a 5-fold cross validation, as it helps prevent the over-fitting (Zheng et al., 2015), by tuning the parameters of a model. Those initial tests are performed with the default parameters set by H2O for each algorithm, as a way to establish a baseline result.

Since the system to be developed needs to generate precise parking availability values, because if the system returns more free parking spaces than there really are (overestimating), it would forward a user to a parking lot with no free parking spaces, but the opposite case is not ideal either, since if the model estimates a value below the available places (underestimating), the system may not refer the user to this park, but to a more distant one revealing a problem for the user and smear the confidence on the system (Richter et al., 2014).

Having said that, the frequency of the outcomes are evaluated, which can be seen in Figures 5.1, 5.2 and 5.3, representing Park 1, Park 2 and Park 3, respectively, as a way to improve the possible prediction results. The values of the park occupancy are shown to be quite unbalanced for all three cases, having always the biggest peak around the 10% mark and the second largest peak, is always around the highest levels of the occupation tax. Using absolute numerical numbers or percentages are not the best solutions, since it can mislead the users, not aware of the total number of parking spaces available (Richter et al., 2014). The occupation rates were placed into interval values, as a way of balancing the results, to help in the perceptibility of the outcome and to increase the accuracy results, thus also reducing possible errors. Taking this into consideration, the results were categorized defined for each parking lot in the next step.

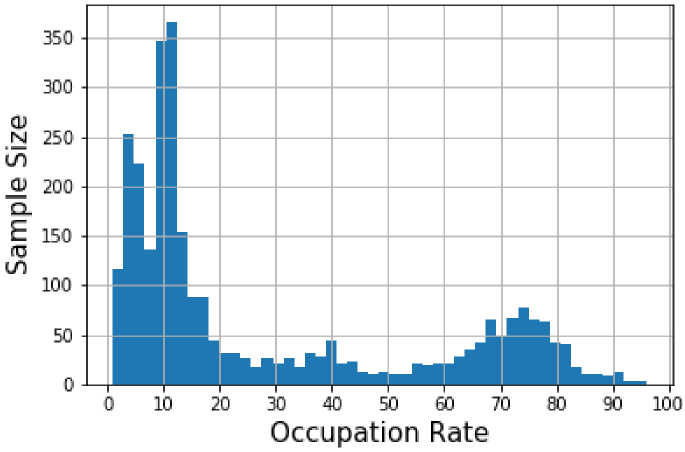


FIGURE 5.1: Histogram values for the occupation rate from Park 1.

Figure 5.1 shows that the most common values for the occupation tax on Park 1 is around the 10% occupancy mark, with also a big frequency on the values smaller than that one, around the values of 5%. A small peak can be seen on the 40% occupancy levels and a bigger peak around the 75% occupancy. These values result on a average occupancy value of 28%. Taking this into consideration, the values were placed into the following categories: 0%-10%, 10%-30%, 30%-50%, 50%-75% and 75%-100%.

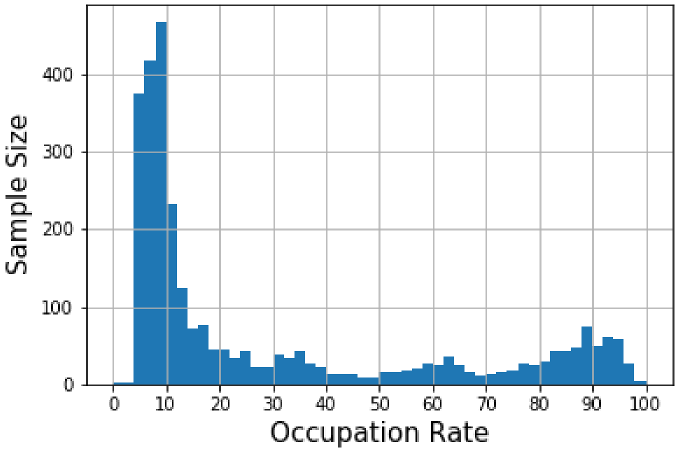


FIGURE 5.2: Histogram values for the tax occupation from Park 2.

The analysis of Figure 5.2 shows that there is, once again, a really high frequency around the 10%, with small speaks on the 35% mark, on the 60% mark and on the

90%, showing an average occupation of 30%. Taking into account this conclusions and by analyzing the Figure 5.2, the values were categorized between the following intervals: 0%-10%, 10%-35%, 35%-60%, 60%-80% and 80%-100%.

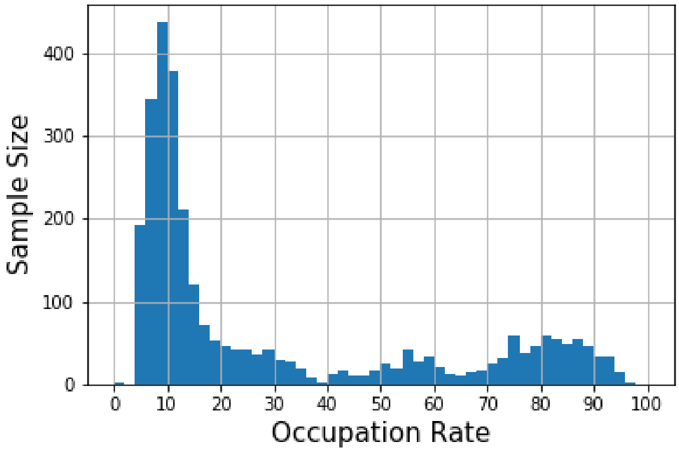


FIGURE 5.3: Histogram values for the tax occupation from Park 3.

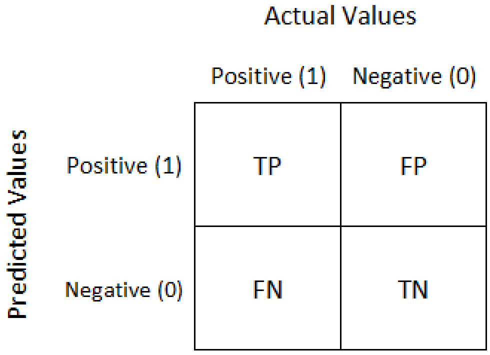
Finally, Figure 5.3 shows that the largest peak occurs at the 10% mark. Other smaller peaks occur, one in the 30% mark, other in the 55% mark, another at the 75% and lastly on the 85%. The average occupation in Park 3 is around the 30% occupation mark. This resulted in the categorization in the following categories: 0%-10%, 10%-20%, 20%-55%, 55%-75% and 75%-100%.

It is important to see that the distribution of the data per category on each parking lot, so in short, for Park 1 the following distribution occurs: the first category (0%-10%) has 989 occurrences, the second and most popular category (10%-30%) with 966 occurrences, the third and smaller category (30%-50%) with 238, the fourth (50%-75%) with 477 and lastly, the fifth category (75%-100%) with 281 occurrences. Park 2 had the following distribution: the first category (0%-10%) was the most common with 1265 occurrences, the second category and second most common category (10%-35%) had a total of 809 occurrences, the third category (35%-60%) with 227, fourth category (35%-60%) is the least common category with only 211 cases and, finally, the fifth category (80%-100%) with 438 occurrences. At last, Park 3 has the following distribution: once again the most common category is the first one (0%-10%) with 1019 occurrences, the second category (10%-20%), being the second most common category, had a total of 783 occurrences, the third category (20%-55%) with 402 cases, fourth category (55%-75%) was the least frequent category with only 241 occurrences and the fifth and last category (75%-100%) with 478 cases.

This categories can still be classified within text categories, as each of the parking lots presents five possible categories, those can be transformed into the following classifications by occupation order, ”empty”, ”low occupation”, ”moderate occupation”, ”high occupation” and ”full”. This categorical division should be implemented according to the need and economic strategy of each parking entity. The categories of the park occupancy allows a more intuitive and perceptible presentation for the user, also leaving a smaller window of error, since even if the park is not within the range of occupation presented, this transformation of the categories does not destabilize the confidence of users.

Real-time system effectiveness depends both on the results and on the time in which these are produced (E. Vlahogianni et al., 2014), so taking this into consideration, the models created are evaluated using the accuracy metric, between the values of 0 and 100, as well as the mean execution time, measured in seconds and being this a solution to be implemented in the application the execution time is important for the application efficiency and speed.

The accuracy metric can be calculated by the evaluation of the confusion matrix that correlates the actual values with the predicted ones. Figure 5.4 shows an example of a confusion matrix, where there is the True Positive (TP), representing the correctly predicted positive values and the False Negative (TN), representing the correctly predicted negative values, also having the False Negative (FN) and the False Positive (FP) which represent the prediction errors. The FN represent the number of negative values predicted as positive and the FP represent the number of positive values predicted as negatives.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

FIGURE 5.4: Confusion Matrix example.

Each of previously referenced components helped to calculate the values of accuracy. The accuracy metric, is the most intuitive performance measure, generally representing the overall performance of the model, calculated using the Equation 5.1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

The execution time consists on the time needed to execute the predictive model, namely the training time plus the scoring time.

5.3 Predictive Model Testing

In this section different models were developed, further leading to some analysis and comments on the results of each model.

5.3.1 Model with full dataset

The model developed in this case is built recurring to the full data elements, concluded at the end of Chapter 4. Following this, the metric results obtained are shown in Table 5.1.

Table 5.1: Accuracy and mean execution time results for the full-data model.

Park	GBM	DRF	NN
Park 1	82% - 8s	81% - 9s	64% - 23s
Park 2	80% - 4s	80% - 9s	69% - 24s
Park 3	80% - 5s	79% - 8s	68% - 23s

Table 5.1 shows that the GBM model reached the best results of accuracy and execution times when comparing with the rest of the models. The GBM model created with full-data reached a maximum accuracy value of 82% with Park 1 data and a total of 80% of accuracy for Park 2 and Park 3. This type of model also showed the lowest execution time values with a total of 8 seconds for Park 1, 4 seconds for Park 2 and 5 seconds for Park 3. The NN created the models with the lowest accuracy levels and the biggest execution times and one of the possible reasons for the accuracy values being so low when comparing to the rest of the methods used, may be because of the small size of the dataset to be used.

5.3.2 Model without context data

In this section a model without any context data (traffic, weather and events data) and only parking data was created, as a way to understand the impact of the contextual data on the prediction accuracy. This model is then built only with the *occupation_rate* column and the date and time features, the results can be seen in Table 5.2.

Table 5.2: Accuracy and mean execution time results for the model without context data.

Park	GBM	DRF	NN
Park 1	80% - 2s	85% - 5s	45% - 20s
Park 2	72% - 2s	78% - 6s	57% - 21s
Park 3	71% - 2s	78% - 6s	49% - 21s

The Table 5.2 shows that the best model in this case was the DRF with a maximum accuracy for Park 1 of 85% and an execution time of 5 seconds. The other two parks showed the same value accuracy and execution time of 78% and 6 seconds, respectively.

5.3.3 Model without December data

A model without the December data was developed as a way to better understand the impact of the occupation rate of this data on the models, since this month presents more irregular patterns. The results obtained in this model can be seen in Table 5.3.

Table 5.3: Accuracy and mean execution time results for the model without December data.

Park	GBM	DRF	NN
Park 1	85% - 4s	85% - 5s	64% - 20s
Park 2	82% - 4s	81% - 5s	69% - 18s
Park 3	79% - 3s	79% - 6s	64% - 18s

Evaluating the results shown by Table 5.3 shows that the GBM model has the best overall results, showing a maximum of 85% accuracy for Park 1 and an average 4 seconds running time. The rest of the parks showed high accuracy levels as well, with Park 2 having 82% and Park 3 with 79%. The DRF model also shows good values of accuracy but with slight increases in the average execution times, where for Park 1 there is a 85% accuracy, but with an higher execution time of 5 seconds. Once again, the NN models showed the worst results for accuracy, but in this case higher than normal, and lower execution times than usual.

5.3.4 Model by month

The following models were built for each of the months being studied, meaning that four models were created for each parking lot, resulting on a total of twelve models. In Table 5.4 the results obtained for each model can be checked, as well as the average results for each parking lot.

Table 5.4: Accuracy and mean execution time results for the models divided by month.

Park	Month	GBM	DRF	NN
Park 1	October	89% - 2s	89% - 2s	76% - 7s
	November	83% - 2s	80% - 2s	70% - 7s
	December	76% - 2s	71% - 2s	57% - 7s
	January	86% - 2s	87% - 2s	69% - 7s
	Average	**84% - 2s**	**82% - 2s**	**68% - 7s**
Park 2	October	80% - 2s	80% - 2s	66% - 7s
	November	83% - 2s	83% - 2s	63% - 7s
	December	78% - 2s	75% - 3s	61% - 7s
	January	78% - 2s	78% - 2s	69% - 8s
	Average	**80% - 2s**	**79% - 2s**	**65% - 7s**
Park 3	October	78% - 1s	76% - 3s	58% - 7s
	November	82% - 2s	79% - 2s	66% - 7s
	December	78% - 2s	72% - 2s	55% - 7s
	January	73% - 1s	72% - 2s	51% - 7s
	Average	**78% - 2s**	**75% - 2s**	**58% - 7s**

Considering the results obtained in Table 5.5, the best result can be seen in the GBM model, for both accuracy and execution time metrics. Overall, the least accurate month is the December month, revealing that the previous assumption is right, in other words the December data is the hardest to the model to learn. In contrast, the months of October and November shows overall good results for the three parks and all algorithms. Having said that, the GBM model shows a maximum accuracy rate for Park 1, once again, with a maximum of 84% accuracy and a low execution time of 2 seconds for each model.

5.3.5 Model by day of the week

For the last test, a model per day of the week for each parking lot was created (resulting on seven models per park, a total of twenty one models). The results can be seen in Table 5.5.

Table 5.5: Accuracy and mean execution time results for the models divided by day of the week.

Park	Weekday	GBM	DRF	NN
Park 1	Monday	72% - 2s	72% - 2s	55% - 5s
	Tuesday	82% - 2s	82% - 2s	69% - 5s
	Wednesday	82% - 2s	82% - 2s	64% - 5s
	Thursday	83% - 2s	83% - 2s	63% - 5s
	Friday	81% - 2s	77% - 2s	59% - 5s
	Saturday	82% - 2s	79% - 2s	67% - 6s
	Sunday	88% - 1s	86% - 1s	65% - 5s
	Average	**81% - 2s**	**80% - 2s**	**63% - 5s**
Park 2	Monday	79% - 1s	79% - 2s	62% - 5s
	Tuesday	81% - 2s	77% - 2s	62% - 5s
	Wednesday	73% - 1s	70% - 2s	61% - 5s
	Thursday	79% - 1s	76% - 2s	65% - 7s
	Friday	84% - 1s	80% - 2s	69% - 6s
	Saturday	78% - 1s	75% - 1s	60% - 4s
	Sunday	90% - 1s	88% - 1s	78% - 5s
	Average	**81% - 1s**	**78% - 2s**	**65% - 5s**
Park 3	Monday	79% - 1s	79% - 1s	56% - 5s
	Tuesday	87% - 1s	83% - 1s	58% - 4s
	Wednesday	76% - 1s	72% - 2s	57% - 5s
	Thursday	77% - 1s	75% - 1s	58% - 4s
	Friday	81% - 1s	79% - 1s	58% - 5s
	Saturday	71% - 1s	69% - 1s	56% - 4s
	Sunday	79% - 1s	77% - 1s	78% - 5s
	Average	**79% - 1s**	**76% - 1s**	**60% - 5s**

Table 5.5 shows once again that overall the most effective algorithm is the GBM, since the models created with this technique show the best accuracy levels and the lowest values of execution time. Park 2 GBM model showed the best values, with an average of 81% accuracy and 1 second of execution time. For Park 1, the same average accuracy can be seen, but a slightly higher execution time of 2 seconds and for Park 3 the same execution time as Park 2 is verified, but a lower accuracy value with 79%. The values from the DRF model are good, not showing a difference of more than 3% in accuracy and the execution time remaining practically the same

when comparing to the GBM model. Once again, the NN model showed the lowest accuracy and the higher execution times. Observing the Table 5.5 shows that the day with the highest accuracy is Sunday, except for Park 3 which is Tuesday. And the least accurate day varies by park, where in Park 1 it is Monday, for Park 2 is Wednesday and at last, for Park 3 is Saturday.

5.3.6 Comparing results

As previously said, the model built with the full-data obtained best results with the GBM algorithm, reaching values of 82% for Park 1 and a execution time of 8 seconds, 80% for Park 2 (with 4 seconds of execution time) and Park 3 (with 5 seconds of execution time).

The model without context data showed overall worst accuracy results, but better execution times when comparing to the full-data model, this is due to the fact that it contains fewer features. In this case, being the DRF model with the best results, reaching an higher 85% accuracy for Park 1 and 78% for the other two.

When comparing the results of the model with full-data with the model without December data, there are better results in the latest one, where this model reached best overall results with the GBM algorithm for both accuracy and execution time metrics. However, these models are not as robust and prepared for outliers cases, as those that are taken into consideration from the month of December, namely because of the events, holidays and vacation periods that happened during this month.

In the case of the models generated for each month, the best results can be seen with the GBM algorithm, where in terms of accuracy there is an increase of 2% for Park 1, but in the case of Park 3 there is a decrease of also 2%, when comparing with the full-data model. However, the execution time has lower values than those obtained with the full-data mode. Despite these improvements, this type of solution requires greater complexity, because it would be necessary to generate eight more models for the missing months of the year, which greatly increased the space occupied in the application, as well as the complexity of adapting the code to deal with various models and the difficulty in re-training of each model.

At last, for the case of the models generated by day of the week the results very similar to those of the model built with the full-data when talking about the accuracy metric, however it got better results in terms of execution time, where Park 1 had an average execution time of 2 seconds and the other two parks with 1 second. However, there is an increase in complexity, as it is necessary to retrain twenty one models in the future and the import of these models into the application does not justify the gain obtained in terms of execution time, as well as an increase in the complexity of the android application.

Having said that, the study continued using the GBM model built using the full-data. Although this is not the model with the best values of accuracy and execution time, it does not depend only on one type of data, meaning that if a problem occurs with one of the sources of data, the efficiency of the model is not fully compromised and the difference between the best accuracy and execution time results to the model built with the full dataset are not considerable.

5.4 Optimization of the Predictive Model

This section focus on optimizing the values of accuracy and execution time for the GBM model built with the full-data, as well as preventing overfitting by adjusting the parameters of the GBM algorithm. For this the focus was on tuning the following options: *ntrees* (specify the number of trees to build), *max_ depth* (specify the maximum tree depth, by default it is 5), *learn_ rate* (specify the learning rate, where range is 0.0 to 1.0.) and *fold_ assignment* (specify the cross-validation fold assignment scheme).

Changing the parameters to the following values, *ntrees* to 50, a *max_ depth* to 7, *learn_ rate* to 0.1 and *fold_ assignment* to ”Modulo”, increased the metrics values, as it can be seen in Table 5.6.

Table 5.6: Models results after optimizing the parameters.

Park	Accuracy	Execution Time
Park 1	85%	5s
Park 2	82%	4s
Park 3	82%	4s

Results shown in Table 5.6 that there was an increase of 3% on the accuracy metric for Park 1 and 2% for the other two and in terms of execution time the value stayed the same for Park 2 and a reduction of 1 second for Park 3, for Park 1 there is a large decrease, now being only 5 seconds of execution time.

To conclude, the confusion matrix of each of the park models was analyzed, that can be seen in Appendix H. The results show that the models have an easier time predicting the values on the first two categories and the latest one, where generally there is lowest error rate. However, the middle classes show the biggest error rates, this is due to the fact that the data are not the most balanced, and despite previous efforts to try to balance the data, these results are still the most complicated to predict. Even so, these models were the ones that continued to be used in the application.

Chapter 6

Decision Algorithm Development

In this section, the main focus is to define the most important heuristics for weight based decision algorithm, as well as the initial development of it.

As it was previously stated, the mobile App option "Navigate" aims to recommend the most optimal parking lot to the driver while taking into consideration various heuristics.

One important factor for the drivers to choose a parking lot is the maximum distance the user is willingly to walk from the park to the final destination (Shin & Jun, 2014) and in (Pullola et al., 2007) the authors establish that this distance increases overtime. So, taking into consideration those requirements, when the user clicks on the ”Navigate” option a new window opens with two input boxes that the user must fill. The first input being the destination where the user wants to travel to and the second input is the maximum walk distance the user is willingly to walk to the final destination. The maximum walk distance the user inserted allows to discard every parking lot outside that radius, meaning that all the parking lots further than that value to the final destination can not be taken into account, so they are automatically discarded. All the other parks inside the valid radius are taken into account to be chosen.

To decide the most optimal parking lot inside the search radius, some heuristics are taken into account, like the duration of route, distance to the final destination and price per hour, since those are the most important factors for the user when deciding where to park (John Golias & Harvatis, 2002). So, taking into consideration those heuristics, a decision algorithm was created based on the Equation 6.1. The equation is based on weight factors that outputs a final weight result, that is used to decide which parking lot is used, since this decision algorithm is applied to all of the parking lots and each parking lot has a weight result associated, the parking lot with the biggest weight value is selected as the appropriated solution.

$$Weight = AR * 0.45 + DR * 0.25 + DPD * 0.2 + PPH * 0.1 \tag{6.1}$$

As it can be seen, the most important factor to take into account for the Equation 6.1 is the Availability Rate (AR) on the parking lot at the time of arrival of the user to the parking lot, measured from the difference between 100 and the occupancy rate of the park. This heuristic has a total weight of 45%, and the occupation rate is obtained by the predictive model, developed in the previous chapter, Chapter 5.

The Directions API (Google, 2019) allows to obtain direction information, namely the time it takes the driver to reach the parking lot, while taking into consideration traffic status. The time needed to reach the parking lot makes it possible to feed the predictive model with the time of arrival and is also used to calculate the Duration of the Route (DR), the second most important factor in the equation, with a weight of 25%.

Next heuristic is the Distance from the Parking lot to the final Destination (DPD) with a weight of 20%, where the closest parking lot to the destination has a bigger probability to be chosen. This information is also obtained through the Directions API, that gives us the distance between two points in meters, namely between the distance between the final destination and the parking lots inside the search radius.

Lastly, there is the Price Per Hour (PPH) of the parking lot, with a 10% influence in the weight result, where the cheapest park has more importance for the weight result. In this case we use the hourly price being this information fixed. The PPH results can be seen in the Table 4.1 for each parking lot.

At the end, the parking lot with the biggest weight value is selected and a route is created using the Google Directions API, starting at the current location of the user to the respective parking lot, directing the driver to that park.

Chapter 7

System Demonstration

In this section, the main focus was the consolidation of the proposed system, namely the import of the optimized model for each parking lot built in the previous section and test its efficiency and functionality within the context of parking availability. An example of the system adapted to the electric charging availability is also shown.

The first step in this section is the importation of the predictive models into the mobile application and this was done through the H20 library (H2O.ai, 2019) which allows to convert the predictive models previously built into a Model ObJect, Optimized (MOJO). H2O-generated MOJO are intended to be easily embeddable in any Java environment.

After importing the models, the mobile application is able to score an occupation category at a specific *datetime* for each parking lot. For the scoring of the predictive model there is a need to get real-time information about the traffic flow and the current weather conditions. So, to get the traffic flow in the surrounding of the parking lots the Traffic Flow API from TomTom (TomTom, 2019) was used, giving information about travel times of the road segment closest to the given coordinates. For the weather conditions, the Current Weather Data API (OpenWeatherData, 2019) was used allowing to obtain real-time information on the weather conditions of the location provided, in this case Lisbon.

7.1 Example A: Parking lot availability

A validation is made in this subsection, namely the testing of the performance of the GBM model and of the decision algorithm with a real case scenario. For this assessment the example used was in Lisbon where a user is trying to find an available place, namely at 9:00 on 17/07/2019 (Wednesday).

For this example, the user started at Instituto Superior de Ciências do Trabalho e da Empresa - Instituto Universitário de Lisboa (ISCTE-IUL) in Lisbon, looking for a place to leave the car in the area of Marquês de Pombal with a radius of 500 metres of search between the destination and the parking lot, where all the parks outside of this range are discarded. If there is not a single option inside the range defined by the user, the application suggests the closest parking lot and a pop up appears asking the user if the application should show the route to that park. In this case, there are three parking lots inside the radius and so, the decision algorithm runs and calculates the value for each heuristic for each of the parking lot, as well as the final weight.

As it been said earlier, this algorithm takes into consideration four heuristics to find the best suitable parking lot. First it takes into account the availability rate of the parking lot at the time of arrival (AR), secondly the duration it takes from the current location of the user to the park (DR), next the distance from the parking lot to the destination provided by the user (DPD), in this case Marquês de Pombal and lastly, the price per hour for having the vehicle parked in the parking lot (PPH). By taking this information, the algorithm executes and provides the result of each option, while also giving the value for the weight result. The option with the biggest value on the weight result is chosen and the route to that parking lot is created.

The results obtained by the decision algorithm, from the example previously presented, can be seen in Table 7.1.

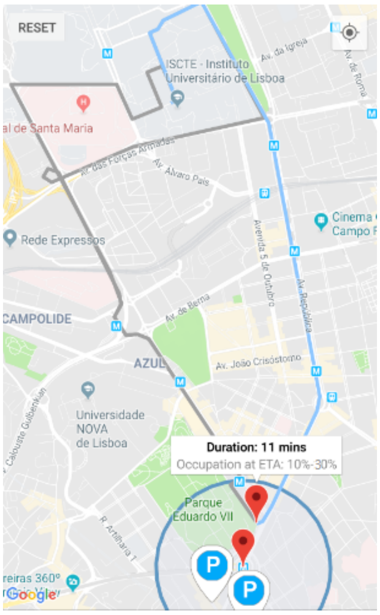
Table 7.1: Decision algorithm heuristics and weight result outcomes for the three parking lots.

Park	Occupation Rate (%)	Duration (Seconds)	Distance to Park (Meters)	Distance (Meters)	Price per Hour (€)	Weight
Park 1	10%-30%	653	262.8	2443.2	1.8	206.84
Park 2	10%-35%	947	347.5	3028.7	2.15	191.60
Park 3	10%-20%	932	273.1	2862.8	2.3	202.78

Table 7.1 shows that all three parking lots have similar occupation rates at the time of arrival in the respective park, being this the most important heuristic to take into consideration by the decision algorithm. At the moment Park 1 has reached the occupation rate of 10%-30%, Park 2 with 10%-35%, and lastly Park 3 with a 10%-20% occupation rate. Next, the second most important heuristic, is the duration of the route to the respective parking lot, and in this case the parking lot with the shortest duration is Park 1 with a route duration of 653 seconds (10 minutes and 53 seconds). Park 3 had the second shortest route duration with a total of 932 seconds (15 minutes and 32 seconds) and finally, Park 2 with the biggest route duration of 947 seconds (15 minutes and 47 seconds). The third most important heuristic is the distance the parking lot is to the final destination (in this case Marquês de Pombal), and by analyzing the table, the closest parking lot is Park 1 with a distance of 262.8 meters, Park 3 shows to be the next closest option with a distance of 273.1 meters and the farthest park is Park 2, with a distance of 347.5 meters. The last heuristic to take into consideration is the price per hour, and the parking lot with the cheapest price is Park 1 with 1.80€, next is Park 2 with a total of 2.15€per hour and lastly, it is the smallest parking lot of all three, Park 3 with a price per hour of 2.30€.

By providing all these heuristics to the decision algorithm it produces an output (weight), which is used for the decision of the optimal parking lot as the best option, and in this case, the parking lot with the highest weight output is Park 1, with a value of 206.84. The other two parks showed a weight value of 202.78 and 191.60, for Park 3 and Park 2, respectively.

In Figure 7.1 the outcome of the decision algorithm can be visualized, has the route to Park 1 is made, having a total occupation at the time of arrival of 10%-30% and a route with an approximate duration of 11 minutes. It can also be verified the maximum search range (off 500 meters) centered in the final destination that it is Marquês de Pombal.



Duration: 11 mins

Occupation at ETA: 10%-30%

FIGURE 7.1: The route to the Park 1 selected by the decision algorithm.

7.2 Example B: Electric charging station availability

The proposed system has the flexibility to adapt to several paradigms. By providing the proper context, the system can be applied to the electric charging stations availability, as it can be seen in (Alface et al., 2019). In this work, the proposed system was adapted to the electric charging paradigm, since electric vehicles tend to park in those spaces. The electric charging stations have specific parking spaces, only allowing the parking of electric vehicles and with the increase of the number of electric vehicles (Pontes, 2019) there is an increasing availability problem. So, the charging sessions datasets from the city of Dundee, Scotland over the period of 1 September, 2017 to 6 September, 2018 (Council, 2019) were used to predict the parking availability in electric vehicle charging stations. One conclusion taken in (Alface et al., 2019) was that the contextual information did not provide any type of improvement for the charging availability results, so only the charging sessions information was used to feed the predictive models. It is also important to say that only three charging stations where present in the system, namely the Queen Street Park charging station, Dundee Ice Arena charging station and for Public Works Department charging station.

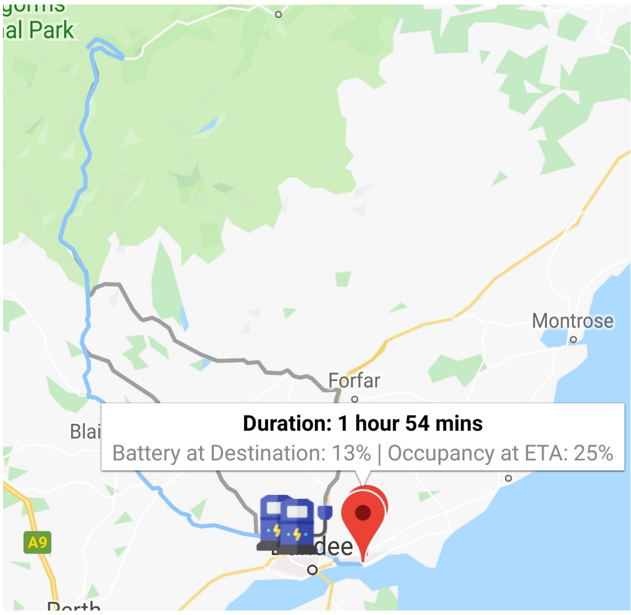
Considering this information, one test was performed with the "Navigate" option to evaluate the efficiency of the system, where the user used a Renault Zoe R90. The journey began at the Braemar, UK, with a 55% battery level, and had Broughty Ferry, UK as the final destination. The results obtained by the decision algorithm can be seen in Table 7.2.

TABLE 7.2: Decision algorithm heuristics and weight outcomes for the three charging stations.

Charging Station	Charging Station Occupation (%)	Duration (Seconds)	Distance to Charging Station (Meters)	Distance (Meters)	Price per Hour (€)	Weight
Queen Street Park	25%	6819	841	68855	0	202.80
Public Works Department	6%	6202	7364	64819	0	201.95
Dundee Ice Arena	0%	6038	9786	63508	0	200.80

The algorithm takes the percentage of occupancy at a higher priority, and the charging station with the lowest occupancy was the Dundee Ice Arena location with 0% occupancy at the time of arrival, as it can be seen by analyzing Table 7.2, while Public Works Department location had a value of 6% and the Queen Street Park location an occupancy of 25%. In the case of the duration, the route with less time was for the Dundee Ice Arena location at a distance of 6038 seconds, about 1 hour and 40 minutes. For the rest of the charging stations, there was a total duration of 6202 seconds (1 hour and 43 minutes) for the Public Works Department location and 6819 seconds (1 hour and 54 minutes) for the Queen Street Park location. The shortest distance from the final destination to the charging station is from the Queen Street Park charging station with only 841 meters away, while the rest of the charging stations were more than 7 kilometers away. In this case, the price per hour was not taken into consideration, since the charging costs are free during the period being evaluated.

So, taking in consideration those values, the decision algorithm generated the weight value, to decide which charging station suits best, and by checking Table 7.2, the charging station with the highest weight value was the Queen Street Park with a total weight of 202.80, where the route to it was generated, as it can be seen in Figure 7.2. The Dundee Ice Arena and the Public Works Department had 201.95 and 200.80 weight value, respectively.



Duration: 1 hour 54 mins
Battery at Destination: 13% | Occupancy at ETA: 25%

FIGURE 7.2: The route to the Queen Street Park charging station chosen from the decision algorithm.

So, it can be concluded that the system proposed, when adjusted to the reality of electric vehicle charging can have good results and can also contribute to a good management and solution to find an electric charging station. In general, this system can also be presented to other parking availability issues, as done to the parking for electric charging.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

The goal of the present work is a development of a proof of concept system based on a mobile application to give information about the parking availability inside the big cities. For this the first step was conceptualizing the desired model, capable of adapting to various paradigms, such as the availability of a parking lot. A system like the one proposed allows the reduction of traffic and pollution within cities, since the driver can better plan his route, or if the driver does not know the city be routed and get information about the available parks, as well as their characteristics. This system presents improvements for the users and for the park management entities, since they are able to direct users to various facilities thus avoiding queues. Another conclusion was that the proposed system can be adapted to the charging station availability, revealing to be a good solution to ease the search for a free charging station.

In this case, the study focused on the parking availability for vehicles from the historical data of three parking lots in the Lisbon area, within a period of 4 months, from October 2018 to January 2019. This data helped to better understand the occupation patterns of the parks, as well as the most important and influential features in occupancy rates between weather, traffic and surrounding events data.

Weather features did not had a great influence on the occupation rate, where the rain and fog occurrence were the only features that showed influence. The lack of correlation between occupation data and weather data could be due to various causes, the first because the parking lots in the study are underground parks, meaning that the weather conditions do not have as much impact than a park on the surface. The second reason is that these parks are highly compounded of covenants vehicles, thus forcing drivers to leave the vehicle in that park, because the cost to park has already been borne by the user. And lastly, the parking lots are categorized as office parking lots, meaning that the area is highly composed of office buildings, and people will have to move, regardless of the weather situation, to their work. In contrast, the months, days, hours, the weekdays, holidays, vacation periods, events in the surroundings and traffic in the vicinity of the parking lot greatly affect the occupation rate in the three parking lots studied. These conclusions can be applied to other parks that are categorised in the same way, in this case categorized as office parking lots.

The choice of the best predictive model assigned to each parking lot was done by building and testing various types of models. For that, different treatments of the data were performed to achieve the best predictor of the occupation rate. The models tested were namely a full-context model (with information about the traffic, events, weather and time periods data), a model only with parking lot data (without any type of context data), a model without December data (since December data is the noisiest and the most complex we take it off to analyze its impact), a model per month and a model per day of the week. Three different algorithms were applied to create the most efficient and accurate predictive model. Those algorithms were the Gradient Boosting Machine, Distributed Random Forest and Neural Networks, evaluated by the accuracy and execution time metrics. Tests show that overall the better models are created using the Gradient Boosting Machine while using the full-context data. This model showed results up to 85% accuracy and with a minimum execution time of 4 seconds.

The models produced and optimized were then imported and incorporated to the mobile application, that was created in order to give the user an interface to interact with the solutions presented in this work. With this and the help of some API services to give us information about the route to reach each parking lot, namely the distance and the duration of the route, as a way to predict at the time of arrival the occupation rate inside the parking lot. This information is then used to decide which parking lot seems to be the most optimal choice for the user with the help of the decision algorithm developed in this work, creating a route to that facility in order to help people who do not know the city or who do not know which is the best option to park the car. The decision algorithm revealed to be fast and efficient to give the best parking lot option to the driver, considering the heuristics used.

Even thought good results and good performance metrics were obtained while using the proposed system, in the next section the discussion on how the proposed solution can be improved.

8.2 Future Work

As said before, there could be some improvements made to the overall system proposed, namely the development of a predictive model to predict the parking availability for on-street spaces, as a way of opening up greater parking possibilities and improving parking management within cities.

It would also be useful to develop a collaborative gamification system to help identify parking spaces along public road, as well as giving the user the option to collaborate with the system by taking a photo of a free parking on-street space, this way the system knows that there is a free parking space, and could be seen in the ”Check Map”. The photo taken to the free on-street parking space needs to be validated, to prove that the photo is taken in the location the user is currently on and that the place is clearly available. If the photo veracity was confirmed users would receive parking discounts, while also improving the efficiency and robustness of the system.

Analyzing other parking lots of different categories, e.g. residential, commercial and other, and incorporating them into the proposed system would also be a great contribution thus opening up options and possibilities for drivers, covering a larger area of the studied location.

Finally, the incorporation of the predictive models into a cloud based server in order to reduce the amount of space needed by the application and increasing the performance of the models, this method would also make it easier to retrain the models, since it would only be necessary to retrain the models on the servers, instead of all the models inside the applications.

References

Alba, E., Chicano, F., & Luque, G. (2017). Predicting Car Park Occupancy Rates in Smart Cities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10268 LNCS*, 107–117. doi: 10.1007/978-3-319-59513-9

Alface, G., Ferreira, J. C., & Pereira, R. (2019). Electric Vehicle Charging Process and Parking Guidance App. *Energies*, *12*(11), 2123. doi: 10.3390/en12112123

Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C., & Vairo, C. (2017). Deep learning for decentralized parking lot occupancy detection. *Expert Systems with Applications*, *72*, 327–334. doi: 10.1016/j.eswa.2016.10.055

Bock, F., Martino, S. D., & Origlia, A. (2017). A 2-Step Approach to Improve Data-driven Parking Availability Predictions. *IWCTS'17 Proceedings of the 10th ACM SIGSPATIAL Workshop on Computational Transportation Science*, 13–18. doi: 10.1145/3151547.3151550

Bock, F., & Sester, M. (2016). Improving Parking Availability Maps using Information from Nearby Roads. *Transportation Research Procedia*, *19*(June), 207–214. doi: 10.1016/j.trpro.2016.12.081

Caicedo, F. (2010). Real-time parking information management to reduce search time , vehicle displacement and emissions. *Transportation Research Part D*, *15*(4), 228–234. doi: 10.1016/j.trd.2010.02.008

Caicedo, F., Blazquez, C., & Miranda, P. (2012). Prediction of parking space availability in real time. *Expert Systems with Applications*, *39*(8), 7281–7290. doi: 10.1016/j.eswa.2012.01.091

Caicedo, F., Robuste, F., & Pita, A. (2006, 01). Parking management and modeling of car park patron behavior in underground facilities. *Transportation Research Record*, *1956*, 60-67. doi: 10.3141/1956-08

Caliskan, M., Barthels, A., Scheuermann, B., & Mauve, M. (2007). Predicting Parking Lot Occupancy in Vehicular Ad Hoc Networks. *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, 277–281. doi: 10.1109/VETECS.2007.69

Chen, B., Pinelli, F., Sinn, M., Botea, A., & Calabrese, F. (2013). Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*(Itsc), 53–58. doi: 10.1109/ITSC.2013.6728210

Council, D. C. (2019). *Electric vehicle charging sessions dundee.* Retrieved from https://data.dundeecity.gov.uk/dataset/ev-charging-data (Last accessed 1 April 2019)

Europe, S. V. (2018). *What is the CRISP-DM methodology?* Retrieved from https://www.sv-europe.com/crisp-dm-methodology/ (Last accessed 31 July 2019)

Gantelet, E., & Lefauconnier, A. (2006). *The time looking for a parking space: Strategies, associated nuisances and stakes of parking management in France* (Tech. Rep.). France: SARECO.

Giuffrè, T., Siniscalchi, S. M., & Tesoriere, G. (2012). A Novel Architecture of Parking Management for Smart Cities. *Procedia - Social and Behavioral Sciences*, *53*, 16–28. doi: 10.1016/j.sbspro.2012.09.856

Google. (2019). *Google maps platform.* Retrieved from https://cloud.google.com/maps-platform/ (Last accessed 28 June 2019)

Greengard, S. (2015, 05). Between the lines. *Communications of the ACM*, *58*, 15-17. doi: 10.1145/2754954

Grodi, R., & Rios-gutierrez, F. (2016). Smart Parking : Parking Occupancy Monitoring and Visualization System for Smart Cities. *SoutheastCon 2016*, 1–5. doi: 10.1109/SECON.2016.7506721

H2O.ai. (2019, 4). H2o documentation [Computer software manual]. Retrieved from http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html (Last accessed 23 April 2019)

Ionita, A., Pomp, A., Cochez, M., Meisen, T., & Decker, S. (2018). Where to Park?: Predicting Free Parking Spots in Unmonitored City Areas. *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, 22:1—-22:12. doi: 10.1145/3227609.3227648

John Golias, G. Y., & Harvatis, M. (2002). Off-street parking choice sensitivity. *Transportation Planning and Technology*, *25*(4), 333-348. doi: 10.1080/0308106022000019620

Klappenecker, A., Lee, H., & Welch, J. L. (2014). Finding available parking spaces made easy. *Ad Hoc Networks*, *12*(1), 243–249. doi: 10.1016/j.adhoc.2012.03.002

Lijbers, J. (2016). *Predicting Parking Lot Occupancy Using Prediction Instrument Development for Complex Domains* (Unpublished doctoral dissertation). University of Twente.

Liu, Q., Lu, H., Zou, B., & Li, Q. (2006, 07). Design and development of parking guidance information system based on web and gis technology. *ITST 2006 - 2006 6th International Conference on ITS Telecommunications, Proceedings*, 1263 - 1266. doi: 10.1109/ITST.2006.288857

Mathur, S., Jin, T., Kasturirangan, N., Chandrasekaran, J., Xue, W., Gruteser, M., & Trappe, W. (2010, 08). Parknet: Drive-by sensing of road-side parking statistics. In (p. 123-136). doi: 10.1145/1814433.1814448

OpenWeatherData. (2019). *Openweatherdata history bulk.* Retrieved from https://openweathermap.org/history-bulk (Last accessed 23 April 2019)

Pflügler, C., Köhn, T., Schreieck, M., Wiesche, M., & Krcmar, H. (2016). Predicting the Availability of Parking Spaces with Publicly Available Data. *Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn*, 361 – 374.

Pontes, J. (2019). *Electric Vehicle Sales Jump 67% In Europe — CleanTechnica EV Sales Report.* Retrieved from https://cleantechnica.com/2019/03/04/electric-vehicle-sales-jump-67-in-europe-cleantechnicas

-europe-ev-sales-report/ (Last accessed 30 July 2019)

Pullola, S., Atrey, P. K., & Saddik, A. E. (2007). Towards an intelligent GPS-based vehicle navigation system for finding street parking lots. *ICSPC 2007 Proceedings - 2007 IEEE International Conference on Signal Processing and Communications*(November), 1251–1254. doi: 10.1109/ICSPC.2007.4728553

Rajabioun, T., Foster, B., & Ioannou, P. (2013). Intelligent parking assist. *2013 21st Mediterranean Conference on Control and Automation, MED 2013 - Conference Proceedings*, 1156–1161. doi: 10.1109/MED.2013.6608866

Ramos Silva, M. H. (2017). *Predicting Space Occupancy for Street Paid Parking* (Unpublished doctoral dissertation). ISCTE-IUL.

Richter, F., Martino, S. D., & Mattfeld, D. C. (2014). Temporal and Spatial Clustering for a Parking Prediction Service. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2014-December*, 278–282. doi: 10.1109/ICTAI.2014.49

Rong, Y., Xu, Z., Yan, R., & Ma, X. (2018). Du-parking: Spatio-temporal big data tells you realtime parking availability. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 646–654. doi: 10.1145/3219819.3219876

Shin, J. H., & Jun, H. B. (2014). A study on smart parking guidance algorithm. *Transportation Research Part C: Emerging Technologies*, *44*, 299–317. doi: 10.1016/j.trc.2014.04.010

Shoup, D. C. (2006). Cruising for parking. *Transport Policy*, *13*(6), 479 - 486. (Parking) doi: https://doi.org/10.1016/j.tranpol.2006.05.005

Stolfi, D. H., Alba, E., & Yao, X. (2019). Can I Park in the City Center? Predicting Car Park Occupancy Rates in Smart Cities. *Journal of Urban Technology*, *0*(0), 1–15. doi: 10.1080/10630732.2019.1586223

Tayade, Y., & Patil, M. D. (2016). Advance Prediction of Parking Space Availability and other facilities for Car parks in Smart Cities. *International Research Journal of Engineering and Technology (IRJET)*, *3*(5), 2225–2228.

Thompson, R. G., Takada, K., & Kobayakawa, S. (2001). Optimisation of parking guidance and information systems display con ® gurations. *Transportation Research Part C: Emerging Technologies*, *9*.

Tiedemann, T., Vögele, T., Krell, M. M., Metzen, J. H., & Kirchner, F. (2015). Concept of a Data Thread Based Parking Space Occupancy Prediction in a Berlin Pilot Region. *Papers from the 2015 AAAI Workshop. Workshop on AI for Transportation (WAIT-2015), January 25-26, Austin, USA*, 58–63.

Tilahun, S. L., & Di Marzo Serugendo, G. (2017). Cooperative multiagent system for parking availability prediction based on time varying dynamic markov chains. *Journal of Advanced Transportation*, *2017*. doi: 10.1155/2017/1760842

TomTom. (2019). *Traffic flow api.* Retrieved from https://developer.tomtom.com/content/traffic-api-explorer (Last accessed 28 June 2019)

Vlahogianni, E., Kepaptsoglou, K., Tsetsos, V., & Karlaftis, M. (2014). Exploiting New Sensor Technologies for Real-Time Parking Prediction in Urban Areas. *Transportation Research Board 93rd Annual Meeting Compendium of Papers*, *14-1673*, 1–19.

Vlahogianni, E. I., Kepaptsoglou, K., Tsetsos, V., & Karlaftis, M. G. (2016). A Real-Time Parking Prediction System for Smart Cities. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *20*(2), 192–204. doi: 10.1080/15472450.2015.1037955

Wang, H., & He, W. (2011). A Reservation-based Smart Parking System. *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 690–695. doi: 10.1109/INFCOMW.2011.5928901

Wang, Z., Yi, J., Liu, J., & Zhang, X. (2007). Study on the control strategy of parking guidance system. *Proceedings - ICSSSM'07: 2007 International Conference on Service Systems and Service Management*, 1–4. doi: 10.1109/ICSSSM.2007.4280300

Xiao, J., Lou, Y., & Frisby, J. (2018). How likely am I to find parking? – A practical model-based framework for predicting parking availability. *Transportation Research Part B: Methodological*, *112*, 19–39. doi: 10.1016/j.trb.2018.04.001

Yang, Z., Liu, H., & Wang, X. (2003, 11). The research on the key technologies for improving efficiency of parking guidance system. In (p. 1177 - 1182 vol.2). doi: 10.1109/ITSC.2003.1252670

Zhang, H., & Li, J. (2018). Deep learning based parking prediction on cloud platform. *Proceedings - 2018 4th International Conference on Big Data Computing and Communications, BIGCOM 2018*, 132–137. doi: 10.1109/BIGCOM.2018.00028

Zheng, Y., Rajasegarar, S., & Leckie, C. (2015). Parking availability prediction for sensor-enabled car parks in smart cities. *2015 IEEE 10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2015*(April), 1–6. doi: 10.1109/ISSNIP.2015.7106902

Ziat, A., Leroy, B., Baskiotis, N., & Denoyer, L. (2016). Joint prediction of road-Traffic and parking occupancy over a city with representation learning. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 725–730. doi: 10.1109/ITSC.2016.7795634

Appendices

A Weather main and weather description values possibilities

TABLE 8.1: Weather main and weather description values possibilities.

weather_main	weather_description
clear	sky is clear
clouds	few clouds (11-24%), scattered clouds (25-50%), broken clouds (51-84%), overcast clouds (85-100%)
drizzle	light intensity drizzle, light intensity drizzle rain, drizzle, heavy intensity rain and drizzle
rain	proximity shower rain, light rain, light intensity shower rain, moderate rain
fog	fog
mist	mist

B Park 1 Occupancy Overtime

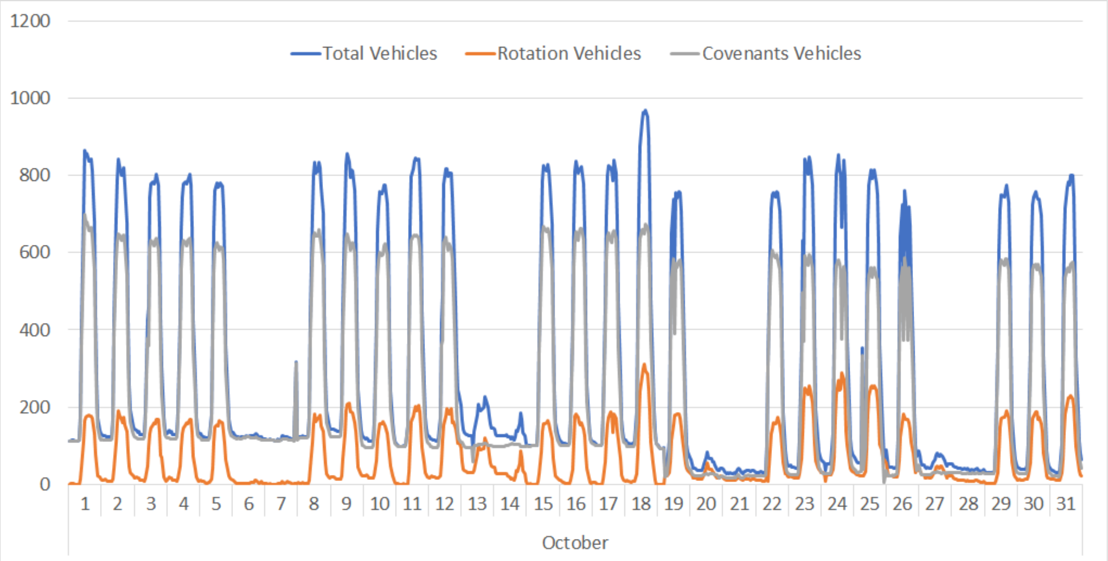


FIGURE 8.1: Park 1 Occupation during the month of October.

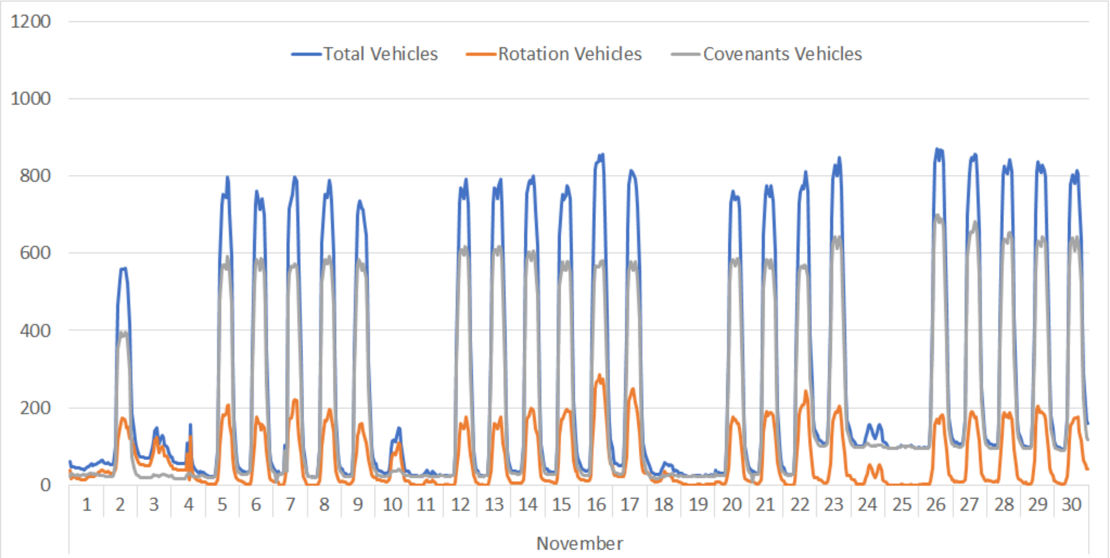


Figure 8.2: Park 1 Occupation during the month of November.

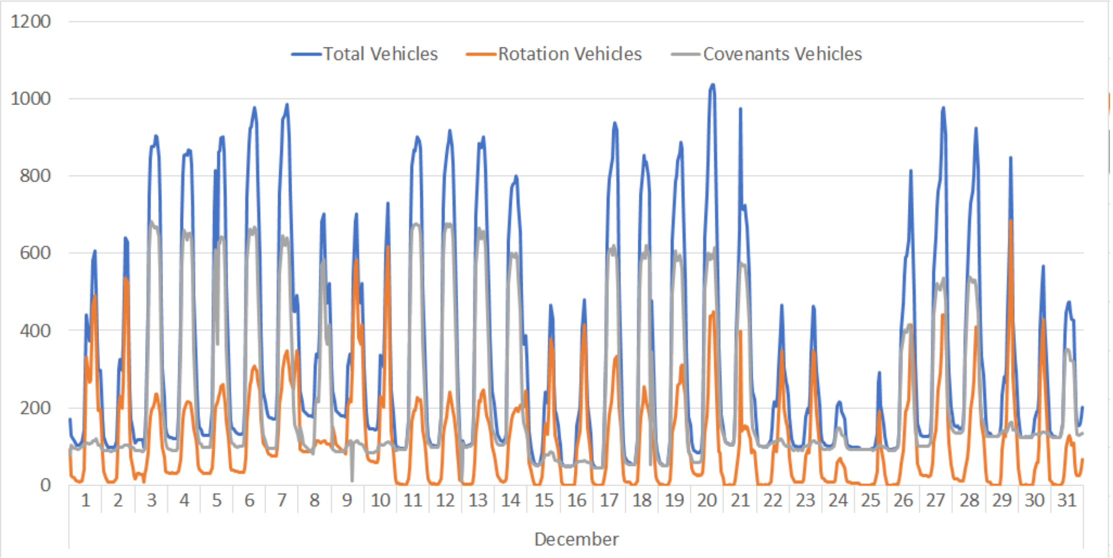


Figure 8.3: Park 1 Occupation during the month of December.

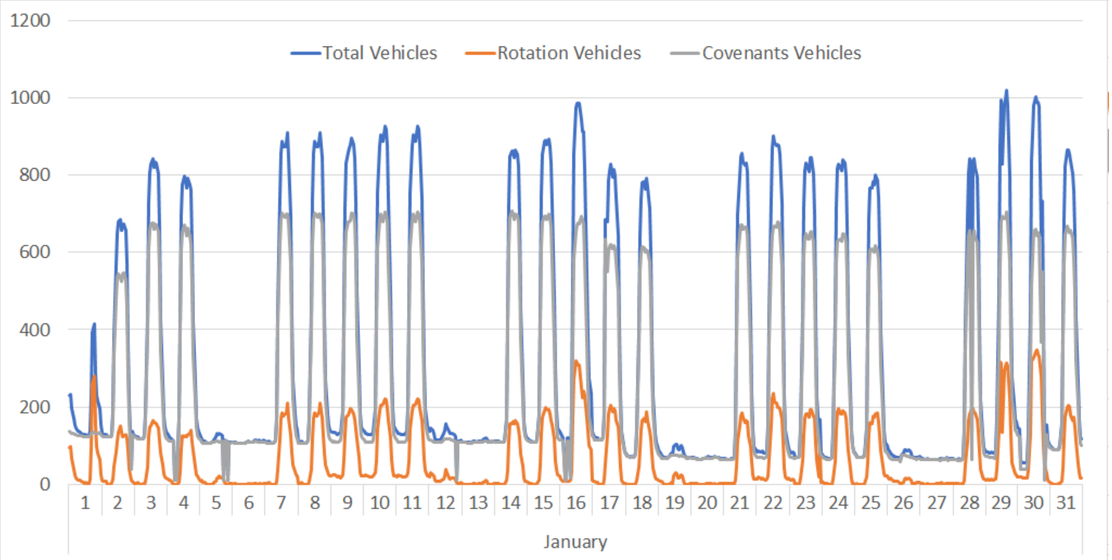


FIGURE 8.4: Park 1 Occupation during the month of January.

C Park 2 Occupancy Overtime

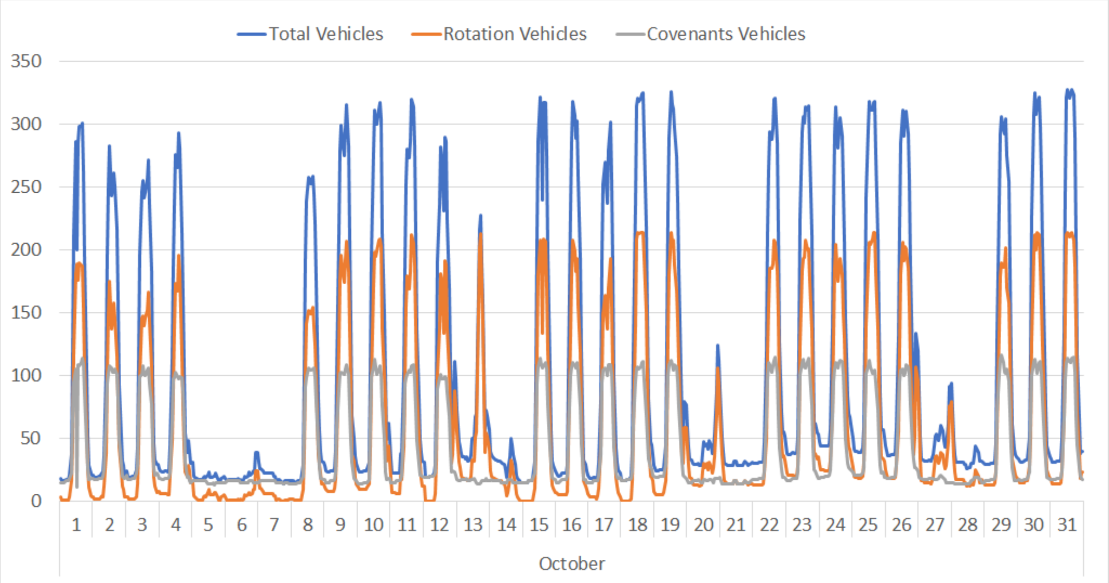


FIGURE 8.5: Park 2 Occupation during the month of October.

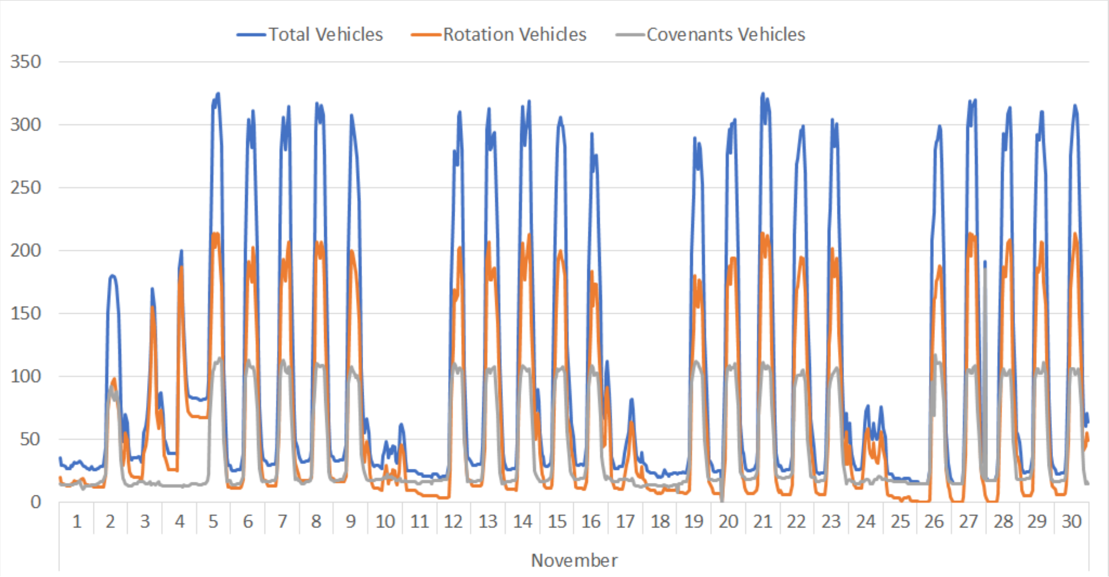


FIGURE 8.6: Park 2 Occupation during the month of November.

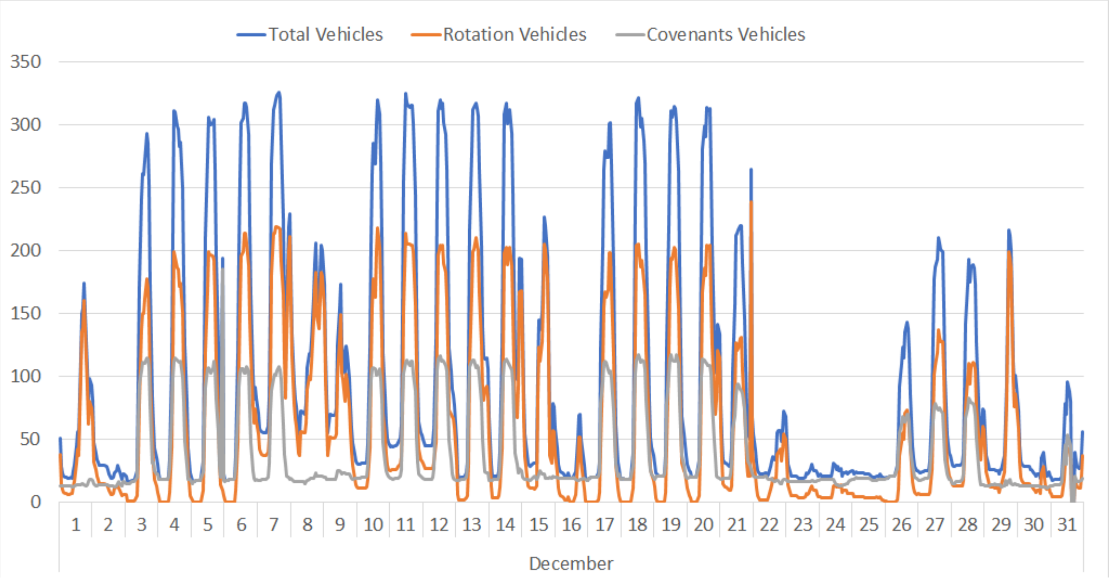


FIGURE 8.7: Park 2 Occupation during the month of December.

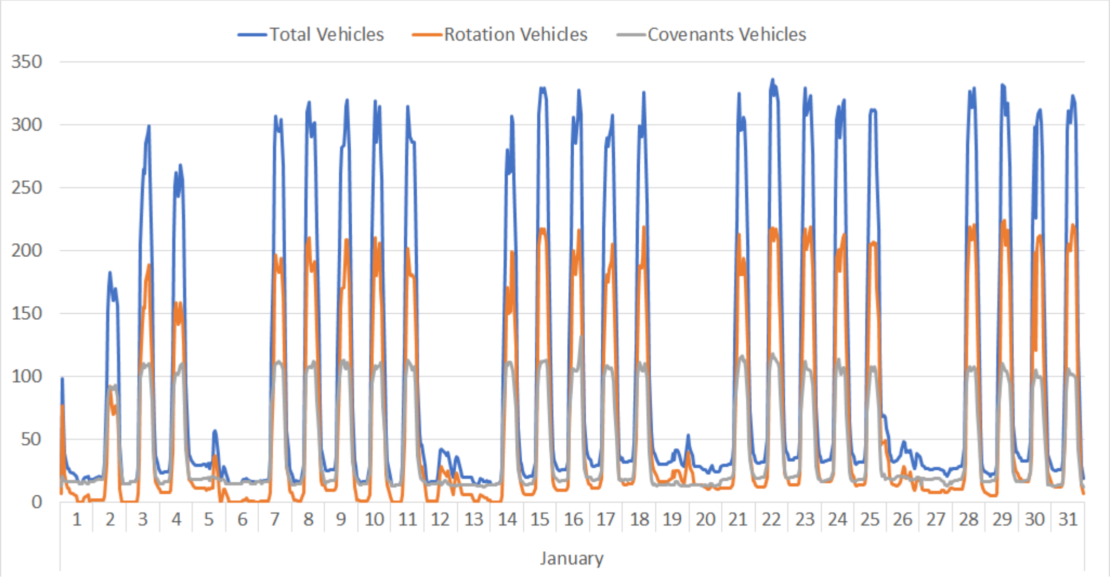


FIGURE 8.8: Park 2 Occupation during the month of January.

D Park 3 Occupancy Overtime

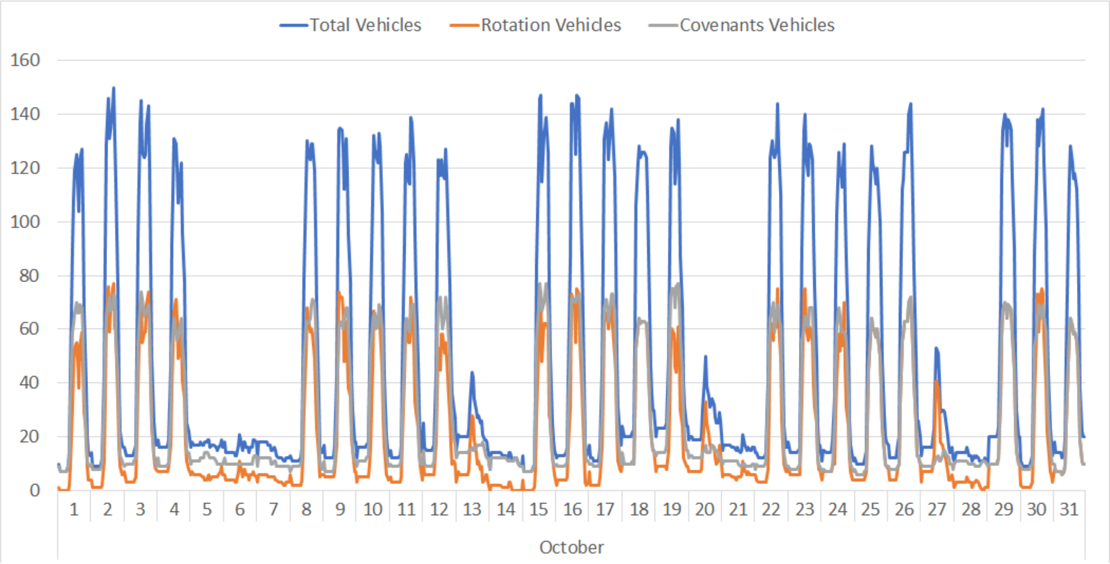


FIGURE 8.9: Park 3 Occupation during the month of October.

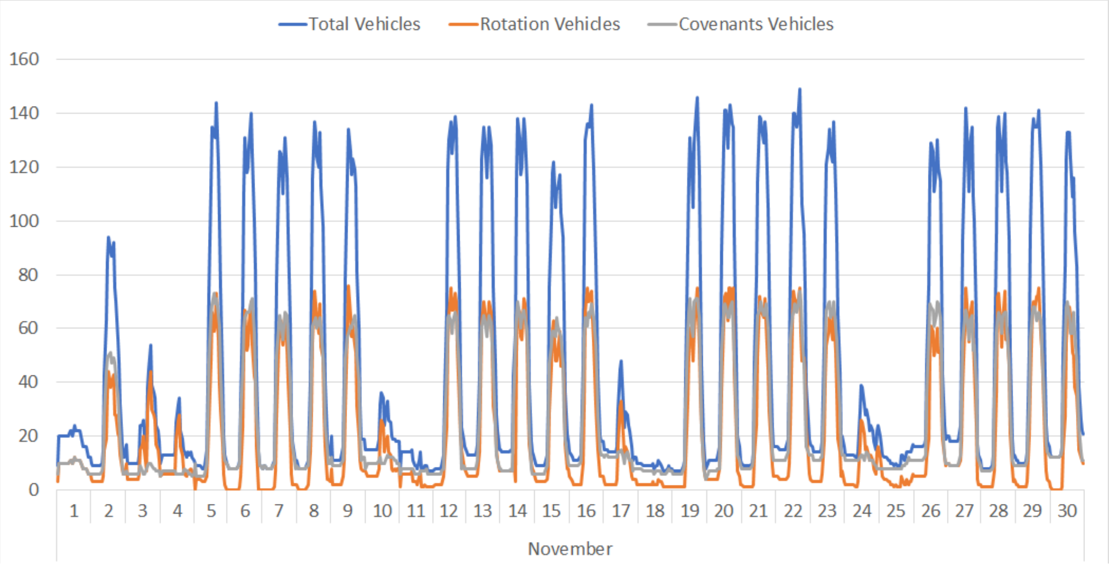


FIGURE 8.10: Park 3 Occupation during the month of November.

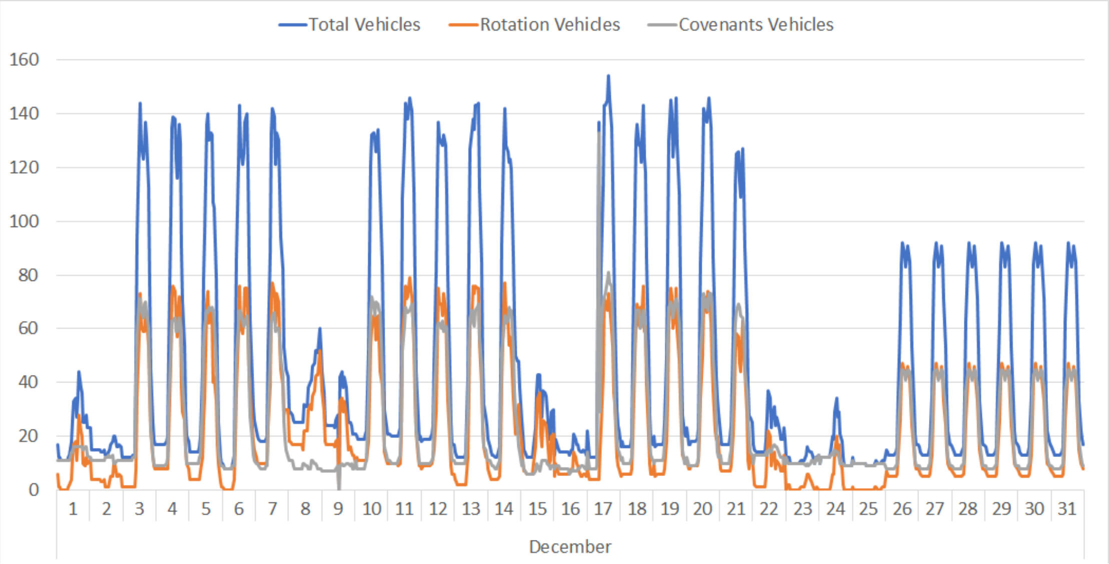


FIGURE 8.11: Park 3 Occupation during the month of December.

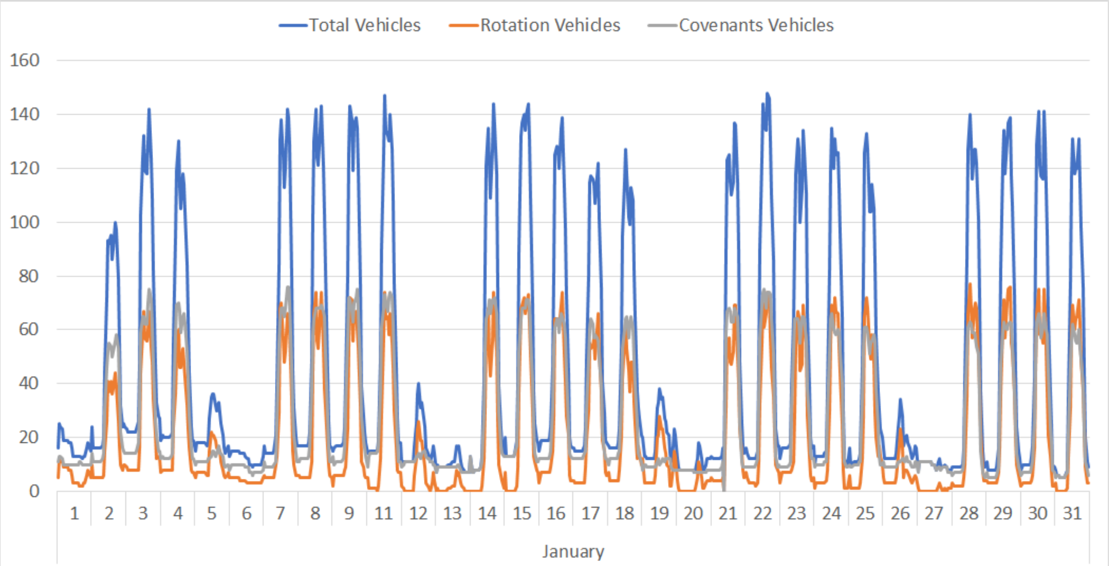


FIGURE 8.12: Park 3 Occupation during the month of January.

E Park 1 Occupancy correlation with Traffic Data

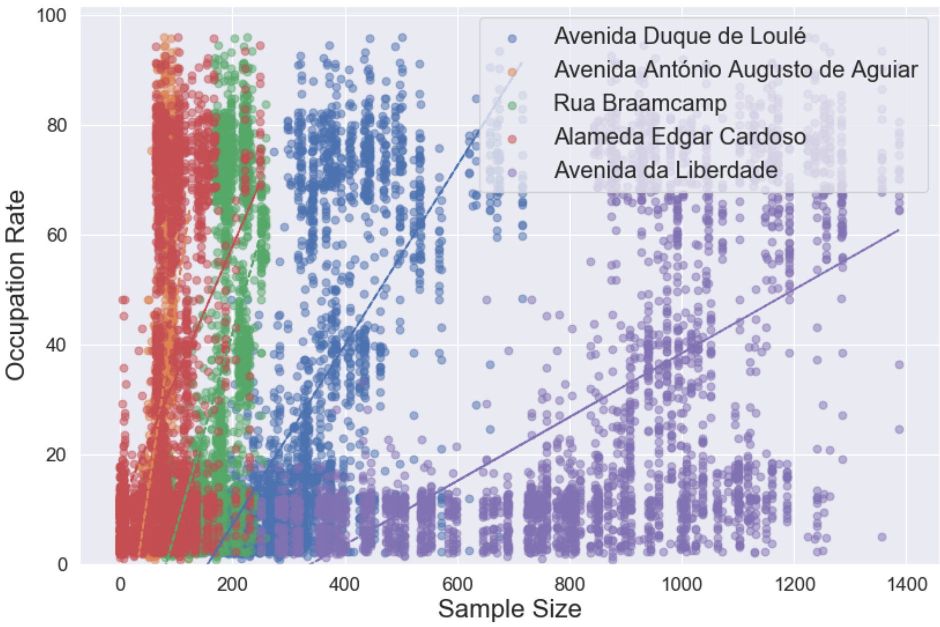


FIGURE 8.13: Correlation between the average time to travel the road segments close to the Park 1 and the occupation rate from the Park 1.

F Park 2 Occupancy correlation with Traffic Data

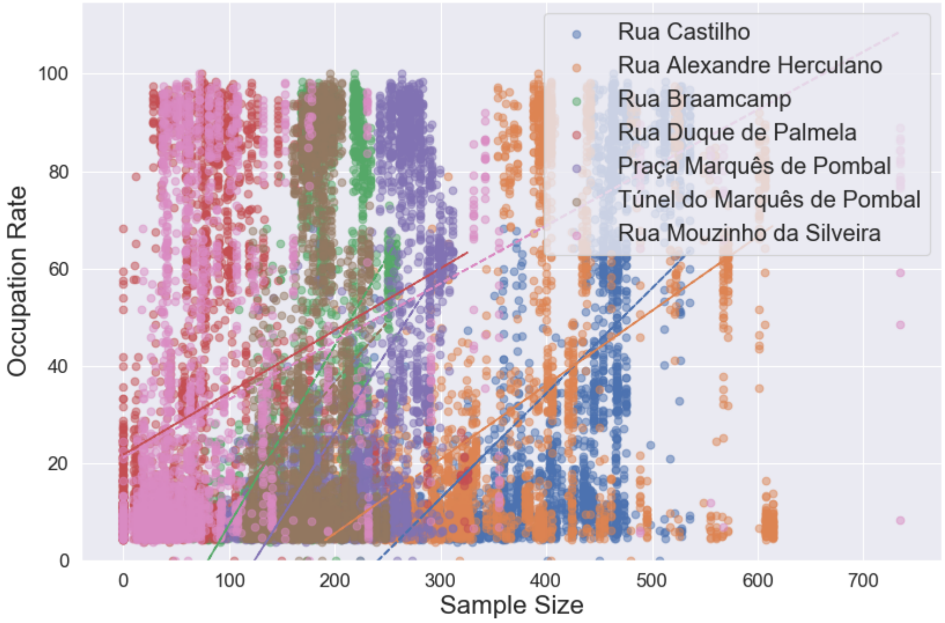


FIGURE 8.14: Correlation between the average time to travel the road segments close to the Park 1 and the occupation rate from the Park 2.

G Park 3 Occupancy correlation with Traffic Data

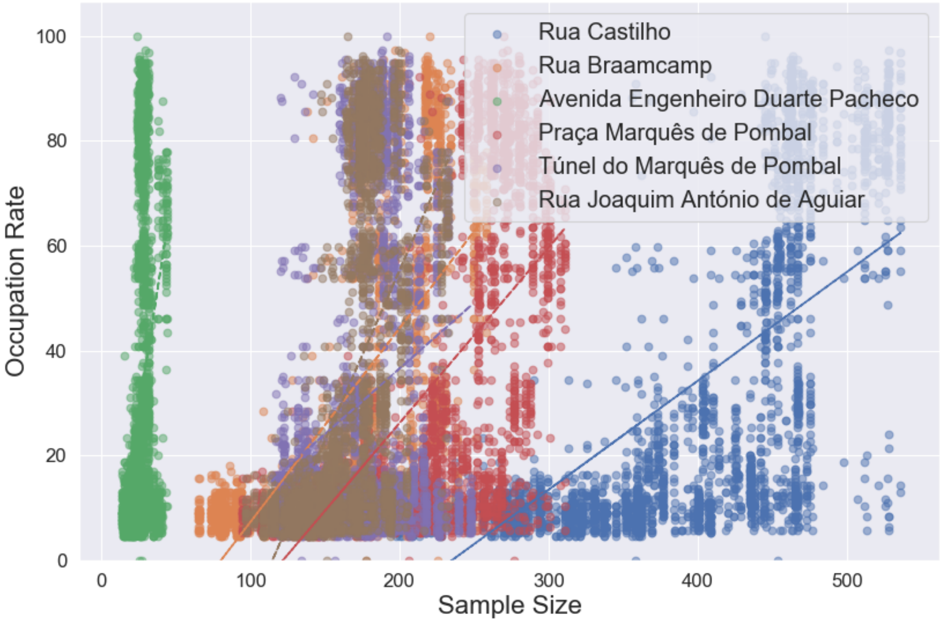


FIGURE 8.15: Correlation between the average time to travel the road segments close to the Park 3 and the occupation rate from the Park 3.

H Confusion Matrix for the Optimized Models

TABLE 8.2: Confusion Matrix of the Optimized Model built with Park 1 data.

0%-10%	10%-30%	30%-50%	50%-75%	75%-100%	Error	Rate
112	16	0	0	0	0.1159	16/138
11	143	0	1	0	0.0774	12/155
1	7	31	2	0	0.2439	10/41
0	1	4	52	9	0.2121	14/66
0	0	1	11	29	0.2927	12/41
134	167	36	66	38	0.1451	64/441

Table 8.3: Confusion Matrix of the Optimized Model built with Park 2 data.

0%-10%	10%-35%	35%-60%	60%-80%	80%-100%	Error	Rate
174	10	0	0	0	0.0543	10/184
21	98	6	0	0	0.216	27/125
0	12	20	7	1	0.5	20/40
0	1	2	19	8	0.3667	11/30
0	0	0	4	58	0.0645	4/62
195	121	28	30	67	0.1633	72/441

Table 8.4: Confusion Matrix of the Optimized Model built with Park 3 data.

0%-10%	10%-20%	20%-55%	55%-75%	75%-100%	Error	Rate
119	19	0	0	0	0.1377	19/138
13	98	9	1	2	0.2033	25/123
1	15	57	4	1	0.2692	21/78
1	0	8	11	19	0.7180	28/39
0	0	0	9	54	0.1429	9/63
134	132	74	25	76	0.2312	102/441