App Guidance for Parking Occupation Prediction

Gonçalo Alface[1], Joao C. Ferreira[1,2](✉), and Ruben Pereira[1]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL,
1649-026 Lisbon, Portugal
{gpaea,jcafa,rfspa}@iscte.pt

[2] INOV INESC Inovação - Instituto de Novas Tecnologias,
1000-029 Lisbon, Portugal

Abstract. This research work presents a prototype model, focused on an android application, to handle the problem of finding an available parking space during driving process for all type of road vehicles in a city using historical data and prediction methods, where there is not any type of real-time system to provide information about the current state of the parking lot. Different source data integration were performed to improve the process of prediction, namely events in the surrounding areas, traffic information on the vicinity of the park and weather conditions on the city of the parking lot. This type of system aims to help users on a daily basis to find an available parking space, such as recommending the best parking lot taking into account some heuristics used by the decision algorithm, and creating a route to it, this way removing some anxiety felt by drivers looking for available spaces.

Keywords: Parking occupancy · Prediction · Mobile App

1 Introduction

Wanting to leave home and knowing that the parking lot will have an empty space for your car when you reach it, is an increasingly necessity in the lives of people. We all have been in the situation of trying to locate a free parking place to park the car, and after minutes and minutes of looking, we start to get frustrate and our stress levels increases, making us angrier and therefore increasing the probability of making an error, possibly causing an accident [6]. The strategy used from most of the drivers looking for a free parking space is called "Blind Search" [14] 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 parking space.

Every day, vehicles in search of free parking spaces negatively impact traffic conditions and the environment, making people lose a lot of time looking for a free parking space. This impact on traffic conditions goes up to 30% [1] and

A. L. Martins et al. (Eds.): INTSYS 2019, LNICST 310, pp. 172–191, 2020.
https://doi.org/10.1007/978-3-030-38822-5_12

pollution in cities for up to 40% [10], as drivers looking for a parking space often slow down or even double-park their cars, which blocks other cars, causing the traffic behind them to slow down too. Another problem lifted with the search for a free parking space, when there is no information, is that drivers often are distracted, putting cyclists and pedestrians in danger [6]. Another emerging problem in the transportation systems in cities is the management of the spatial resource since it is limited, as well as the parking cost being to expensive [14]. This limited resource contributes to cars spending to much time looking for parking places and consume to much gas/energy during the search for an available parking space.

Cars where initially invented to increase convenience and comfort in everyday life of people, however car congestion in a city causes unpleasant problems such as environment issues, energy consumption, parking space shortage, traffic jams, noise, air pollution, and even minor psychological damage to some people [14]. From all of those, we can check that the parking space shortage is 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, and as a result, cars will need more time and have a larger energy consumption while looking for a parking space. A study on the parking situation in Schwabing (Germany) was done and it 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 [2].

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 [18]. On average it takes 12 min to a driver to find a free parking place [10] and a nation-wide 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 [1]. 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 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 L of gas wasted and a total of 730 tons of carbon dioxide produced in one year [7].

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 [16], 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 [13]. Predictive parking information reveals to be a very useful information 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.

Taking all this into consideration, in this paper we propose a system that offers the user the shortest path to the most optimal parking lot considering various conditions, namely the distance from the parking lot to the destination the user wants to get to, the duration of the trip, the occupancy of the parking lot at the time of arrival and the price per hour of parking the vehicle in the park. Information about the parking availability on the parking lot at a respective time must be predicted by taking into consideration historical data from the parking data occupancy, as well as other external factors, like weather, events and traffic and characteristics of the parking lot, like the price per hour and the total number of parking spaces.

This paper will be organized as follows. The following section will be about previous research works for parking lot availability prediction. In Sect. 3 the conceptual model of the proposed system is explained. For Sect. 4 the predictive model is developed and feature selection is made. Next, in Sect. 5 validation for the proposed system is made. Finally, we conclude our paper and suggest future work in Sect. 6.

2 Literature Review

In this section we will focus which features show to be more important and have a bigger impact on the prediction of the park availability, as well as the type of models created to deal with this type of problem.

2.1 Feature Selection for Parking Prediction

Systems based on historical available data are cost-effective and, 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 [15]. In [16] six months of historical data was used, and in [10] 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.

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. Monitoring each single parking space could reveal expensive, so monitoring the flow of entering and leaving the parking lot [7] it is easier and this way the monitor park 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.

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 [8]. Bad weather conditions could lead to lower traffic flow than expected, but parking occupancy would just be affected in shopping malls, iconic locations, and other, not on parks close to apartments and offices [13]. The period of the day and time of year are also important [18], as holidays, weekdays and hour of the day could have direct impact on park occupancy. 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 park close to apartments and shopping malls, and quiet less in office parks [13]. The time of the day can generate more information like, the day, the month and the hours of the day, affecting, once again, the traffic and park occupancy [10]. The location of the parking lot is also an important factor [10], since 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 football games and concerts, those can cause a significant increase in the amount of traffic, consequently increasing the demand for free parking spaces [6]. In [13] each parking lot is categorize into seven categories of the parking lot, being those, apartment, office, mall, food, hospital, park and entertainment. If the parking lot is close to shopping centers or supper markets, it 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.

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 influence the parking occupancy [16].

For [14] 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 parking guidance.

When predicting parking availability, factors like spatial and temporal have varying importance [13], so first there must be a 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 [10].

2.2 Modelling for Parking Occupancy Prediction

In [16] data obtained wirelessly from a IoT sensor network available in the "smart" city of Santader, Spain, giving the current status of the parking space (free/occupied). The model was developed using a methodology of two modules, the first using Neural Network (NN) for the prediction of the time series of parking occupancy in different regions of an urban network, and the second module using some factors (e.g. weekday, weekend, time period, morning evening) with survival analysis for estimating the probability of finding a free parking space in the following time interval, resulting on visual representations, to help the user decide. A naïve prediction is used as baseline, proving that the model generated as better accuracy values than the baseline when the predictive horizon gets larger, revealing a robustness of the NN model in dealing with problems of ranging levels of complexity up to half an hour prediction ahead.

For the development of the application Du-Parking, Recurring Neural Network (RNN) with the incorporation of Long Short-Term Memory (LSTM) were used [13]. RNN have been successfully applied to sequence learning tasks, and with help of LSTM, RNN will be able to continue to learn long-term temporal dependency. For performance analysis, the model built was compared with other two methods, namely Linear Interpolation being a distance-weighted interpolation algorithm with the idea that as the distance between parking lots, its parking availability is similar. The other method used was Gradient Boosting Decision Tree GBDT, and the reason to choose this algorithm goes from its effectiveness on training and on classification. The results of this experiment show that GBDT outperforms Linear Interpolation, and in the case of the Du-Parking model it gives a bigger improvement over the baseline algorithms implemented.

The idea behind [6] study is creating a demand profile, reflecting the parking occupancy in a determined time and area, with the idea to apply this profile to other areas with similar conditions. 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. A step to take into account when implementing a system like this in a new area, is that the data needs to be coherent with the new location. Aggregated features like date and time, traffic value, temperature, precipitation, payment type and payed amount, under a location unit id, are used to train the model and the target variable is the occupancy rate. With the use of similar functions like cosine similarity and earth mover's distance, the authors were able to compare different locations between booth projects, establishing which models best adapt to the areas of each other. Algorithms used on model training using SFpark data, applying algorithms like Decision Trees, Support Vector Machine (SVM), Multilayer Perceptrons and Gradient Boosted Trees. Extreme Gradient Boosting was had the best accuracy result of all of them.

For [18] the modelling of the occupancy rate is done after applying different features with the help of methods like Regression Trees, SVR and NN. Predictions are made for periods of 15 min ahead and all three algorithms were used, but for predictions higher than 15 min, SVR was not, due to the long computation time needed. The first feature set has the input of time and day of the week and the second feature set has the input previous observations of the occupancy rate on a certain time and the number of steps ahead to be predicted (each step represents 15 min). After applying all algorithms results show that Regression Trees, the least computationally intensive algorithm from all three, is better when comparing with the NN and SVR, for all feature sets. The feature set that reveals better results is the feature set that includes the history of the occupancy rates along the time of the day and the day of the week.

As a way to deal with the parking problem and to cope with the limitations from Parking Guidance and Information System, the study in [14] focus on creating a concept of smart parking guidance system, as well as a parking guidance algorithm to assign the driver to the most appropriate parking facility considering various factors, like 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. This information is then evaluated by the parking guidance algorithm in the central server, suggesting 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 he/she arrives, and subsequently parking costs occurs from this point on, or just drives to the suggested option without a reservation.

3 Conceptual Model

In this section we show the conceptual model of the system being developed. As we can see in Fig. 1, the central part of this system is the android application where the user can interact with the system. This application integrates all services present in the study, having access to the predictive model being developed, which are exploited in Sect. 4, with the help of the complete dataset, that combines weather, traffic, events and parking occupancy data, being described in this section. Other services are necessary for the development and proper functioning of the application, namely some Google API Services, like Directions API, Maps API and Firebase. In the following sections we explain the components of the proposed solution being created.

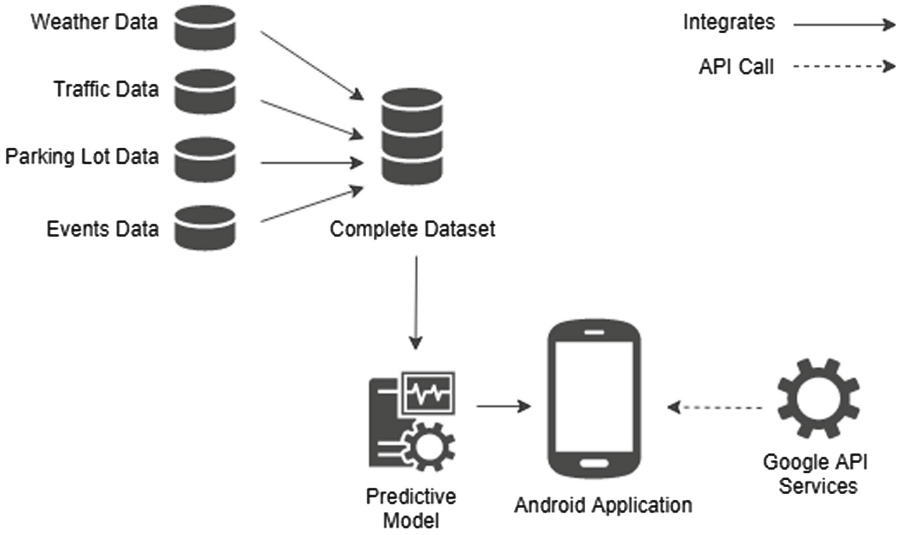


Fig. 1. Conceptual model diagram for the proposed system.

3.1 Android Application

The android application, as have been said, integrates all of the previous services taking into account the predictive model responsible to predict the available parking places in a respective parking lot. The main idea of the application is to contain information about various parking lots so that each one has associated a predictive model that provides the occupancy rate of the parking lot at a given time. By having several parking lots added, this allows a better management of the occupation of these parks, not allowing one park to have high occupancy and the others to be empty, but rather indicating better parking options to the user taking into account a decision algorithm that takes weight dimensions such as, distance from the parking lot to the final destination, duration of the trip to the park, the hourly price of the park, the distance the user is willingly to walk, and more importantly, the occupancy rate on arrival to the parking lot considering the current position of the user. With those heuristics in mind, the decision algorithm provides the most optimal parking lot for the driver. In this case, we focused on a single park, but the robustness of the model allows it to be implemented in more than one nearby parking lot.

The application firstly shows some options that the user needs to fill, namely the destination and how many meters it would like to walk from the parking lot to the final destination. With the all options filled up, the application creates the most optimal route between the start location and the closest parking lot to the end location, by running the decision algorithm, as well as taking into account the range off how many meters the user would like to walk from the parking lot. In [11], if no park is in the surroundings of the end location inside the range defined by the user, the system gives information to the user about which parking lot in the surroundings of the destination has the most probability of having a free parking space, as we have done, while also providing information about the distance from the final destination.

3.2 Complete Dataset

The complete dataset is used to create the predictive model. This dataset is composed by four datasets combined during the period from 1 of October of 2018 to 31 of January of 2019, making it a total 4 months of data. The data was gathered every hour making a total of 4 months of data, resulting on a 2952 row dataset. The first dataset being used is the parking occupancy data from a parking lot in Lisbon situated around the Marquês de Pombal area, plus weather data of Lisbon, events on the surroundings of the parking lot and the traffic data surrounding the parking lot. This data is then merged by the date and time each measurement was made for each dataset. In the following sections we explore each dataset and its composition.

3.3 Parking Lot Dataset

At first we have the parking lot dataset being used in this study, this parking lot has a total of 336 parking spaces and is open 24 h a day from Monday to Sunday.

Even though the dataset as not an extensive size to see the annual pattern of the parking lot, this intercepts a key moment for park affluence, namely during the Christmas period. This period allow us to analyze and perceive how the parking occupancy changes during festive periods, as holidays can have an huge impact on the parking lots [3]. Analysing this dataset we were able to conclude that the months of October, November and January show very similar patterns of occupancy over time, yet the month of December shows a very high variety, much because it is a festive month and because it contains several holidays. Another important factor is that this parking lot is located on a area which is surrounded by office buildings, so we categorize this parking lot as an office parking lot which may prove important in terms of their affluence and time periods [13]. A decisive characteristic for the parking lot occupancy is that it has an associated cost per hour [14] that will be further explain. The following information represents the composition of each row of the dataset:

- Hour - representing the time the measurement was made;
- Date - representing the date the measurement was made;
- Rotation - number of rotation cars inside the parking lot;
- Covenants - number of covenants cars inside the parking lot.

The measurements of Rotation and Covenants gives us the total of cars inside the parking lot from each type of client, and by adding the two we can have the complete parking occupancy in the parking lot for the determined date and hour.

The parking lot has two types of clients, the ones from rotation and from covenants. The rotation vehicles are the ones that enter and leave the parking lot without any kind of commitment, besides having to pay the ticket for the total number of hours spent in the car park. Users of type rotation need to pay a fee per hour, more specifically 2.15€. In the case of the covenants, where a use has unlimited entry and exit from the park during the time period in which he made the advance payment, the cost must be negotiated depending on external factors.

It is important to note that the parking lot to be studied is an underground park with a total of 336 parking spaces and is equipped with extra services like CCTV, so the security is higher, WC for the drivers and passengers, parking places for people with reduced mobility and car wash, and can therefore be differentiating factors at the time the users decide which parking lot they should park the car. The location of the parking lots also turns out to be quite important at the moment of decision by the users [4], since a good location can define the use of a park.

3.4 Weather Dataset

The weather dataset was obtained from the OpenWeatherMap using the API service History Bulk [9] from 1 of October of 2012 to 14 of March of 2019. This dataset comes with 50947 rows, representing a total of almost 6 and a half years of weather data from Lisbon, where each row of the dataset represents a measurement done for a respective date and time. The weather dataset was collected in intervals of 1 h, as the weather conditions typically do not change much during short time horizons [3]. From this dataset we used the following columns:

- dt_iso - date and time in UTC format;
- temp - current temperature in Kelvin;
- temp_min - minimum temperature at the moment in kelvin;
- temp_max - maximum temperature at the moment in kelvin;
- pressure - atmospheric pressure (on the sea level) in hPa;
- humidity - humidity in %;
- wind_speed - wind speed measured in meter per second;
- clouds_all - cloudiness in %;
- weather_main - group of weather parameters (Rain, Snow, Extreme etc.);
- weather_description - weather condition within the group.

Weather_main represent the weather condition within the following categorizations: clear, clouds, drizzle, fog, haze, mist, rain, smoke, snow and thunderstorm. The weather_description gives some more information within the weather_main condition, like if it raining heavily.

It is important to know that temp_min and temp_max are deviations from the current temperature that is possible to happen in large cities and megalopolises geographically expanded.

We transformed some columns to a unit that is clearer and simpler to analyze, such as passing the columns temp, temp_max and temp_min from the Kelvin unit to Celsius. In the case of wind_speed, we decided to convert it from meters per second to kilometers per hour.

3.5 Traffic Dataset

Other type of data used for enrichment of the analysis were the traffic data. 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 influence the parking occupancy [16]. This data gives us information about the traffic state on certain roads and it was obtained the same period as the parking lot data. The data was gathered for the surroundings of the parking lot since those areas in Lisbon are heavily influenced by traffic, and came in the format of a JavaScript Object Notation (JSON), with various interesting components like the average speed, average travel time and speed limit on certain roads, but in our case we opted to use the sample size values that gives us the total number of vehicles that passed through the segment of the road in the respective date and time.

3.6 Events Dataset

One important factor that may over-saturate the parking lot are the events on the surroundings of the location of the park [17]. If rich historical 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. For the events dataset case, we collect all major events that occur within the surroundings of the parking lots for the same data period. Having said that, 13 events were referenced due to their large size and proximity to the parking lot under study, namely the following ones: SIL - Salão Imobiliário de Portugal (03/10/2018–07/10/2018), Web Summit (05/10/2018–08/10/2018), Lisbon Fashion Week (11/10/2018–14/10/2018), Greenfest (11/10/2018–14/10/2018), Lisbon Marathon'18 (14/10/2018), Lisboa Games Week (15/10/2018–18/10/2018), Doclisboa (18/10/2018–28/10/2018), Concert: Kodaline (24/10/2018), Super Bock em Stock Festival'18 (23/11/2018–24/11/2018), Feira Outlet (07/12/2018–09/12/2018), São Silvestre de Lisboa Race'18 (29/12/2018), Wonderland Lisboa (01/12/2018–01/01/2019) and Beethoven's Violin Concert (18/01/2019–19/01/2019).

We did the same to define the holidays, resulting on total of 10 holidays with different importance, meaning that for the same time span of the parking lot data not all of the holidays have the same importance. We decided to give the holidays an importance between the range 0 to 2, where 0 is a normal day, as those do not represent any type of public holiday, value 1 represents a festive day, but it's not officially a holiday, and 2 a very important one. For example, Christmas Day has a bigger impact then the Kings' Day, as we can see in Table 1.

Table 1. Public holidays importance inside the period being analyzed.

Date	Public holiday	Importance
05/10/2018	Implementation of the Republic	2
01/11/2018	All Saints Day/Bread Day by God	2
01/12/2018	Restoration of Independence	2
08/12/2018	Immaculate Conception Day	2
25/12/2018	Christmas Day	2
26/12/2018	Boxing Day	1
31/12/2018	Réveillon	1
01/01/2019	New Year's Day	2
06/01/2019	Kings Day	1

We also defined a school vacation period between 14 December 2018 to 31 January 2019 and a work vacation period between 22 December 2018 to 1 January 2019.

3.7 Google Services

The Google API services are of great importance in this application, as the entire application works around the maps service provided by Google. The Maps API allows the presence of a map based on Google Maps data, that the user can explore by the route to the parking lot. Another service used in the application is the Directions API, that allows the application to obtain direction information and draw a route between two points, taking into consideration traffic stats. This service provides us with information for different transport modes, waypoints and travel times, as well as one or more travelling routes to the destination while checking the distance and time it takes from one point to another. In this application, the only considered transport mode is a car, since all of the studies focus on the parking lot state. Other Google service used is Firebase that supplies the means to build an authentication system for the login and register option, as well as the database to keep information about the users and the parking lots.

4 Development of the Predictive Model

In this section we be developed the predictive model while taking into consideration the different features previously referenced, as a way to understand which features best help the prediction problem reach better accuracy values.

4.1 Data Analysis and Processing

In this step we initially started by combining all the previously referenced datasets by the *datetime* column all of them have and processing and combining some data so it would be easier to analyze it and help take conclusions.

We then did a quick analysis of the data and we came across two incorrect parking occupation measurement, where the total number of cars was greater than the maximum capacity of the car park, which is 336. Having said that, we removed the measurement with values greater than this capacity.

With the help of the *datetime* column, 4 new columns were created, namely year, month, day and hour, representing the hour and the measurement was made, respectively, and after that we dropped the *datetime* column.

To verify which features suits best to predict the total occupation in the parking lot, we firstly created the column *total* that is calculated by joining the rotation values with the covenants values to gives us the total number of vehicles in the car park. We then used the Pearson's Correlation Coefficients, as this method give us a way to establish the relationship between two values, by indicating that if one variable changes in value the other variable tends to change its value too, in a specific direction. The Pearson's Correlation uses two metrics to evaluate the relationship between two variables, namely the strength and the direction. The strength metric makes it possible to understand the absolute correlation between two variables, so the stronger is the relationship, the higher is the number in absolutely from a range between -1 and 1. A relationship of 0 means that there is no type of relationship between two variables is meaningful, and in contrast, if the coefficient is 1 that reveals a really big correlation. The other metric used is the direction of the correlation, namely the sign of the value, where negative values means that, when one variable increases the other

Table 2. Correlation of weather features with the target total.

Feature	Correlation coefficient
road_tunel_marques	0.57
road_mouzinho	0.74
road_duque	0.39
road_castilho	0.63
road_braamcamp	0.55
road_herculano	0.54
road_praca_marques	0.63
hour	0.28
day	0.01
month	0.004
year	0.0075
humidity	−0.24
wind_speed	−0.17
clouds_all	0.11
pressure	0.04
temp_max	0.36
temp_min	0.29
temp	0.32

decreases, and when the value is positive, as one variable increases the other increases as well.

We then applied the Pearson's Correlation to the continuous variables, as this method evaluates the linear relationship between two continuous variables, obtaining the results shown in Table 2.

As we can see from the results shown in the Table 2, the most meaningful features are the traffic data, reaching values of 0.74 and not lower than 0.39. In terms of weather data, we can conclude that only the temperature features show a high correlation values which may be due to the fact that the car park is categorize as an office park [13]. The resulting column features resulting from the *datetime* feature, namely the year, month, day and hour, only the hour feature revealed a bigger correlation value with a total of 0.28, the other features reveal real poor correlation values, no greater than 0.004 in the case of month feature.

Having said, from the features evaluated by the Pearson's Correlation we selected the following columns for the rest of the work: *road_tunel_marques*, *road_mouzinho*, *road_duque*, *road_castilho*, *road_braamcamp*, *road_herculano*, *road_praca_marques*, *year*, *month*, *day*, *hour*, *temp_max*, *temp_min* and *temp_min*. Even though the*year*, *month* and *day* features had low values of correlation, this values may be needed to retrain the model with data from other months and years, just like in [10].

Next, as a way to have better predictive results we created a column named *occupation_tax* that divides the total column by the total park capacity. This method results in better predictions and a more intuitive result to show to the user, meaning that for the case of a total of 280 cars in the parking lot, it represents a 83% occupation that we always round up to multiples of 10, resulting on a total of occupation of 90%, as 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, 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 [12] so we always round up to prevent those types of problems. We then removed the covenants, rotation and total columns, as those columns are highly dependent and highly correlated with the *occupation_tax* target.

Various flag columns were added, the first one was *flag_weekend* that identifies if the current day of the measurement is on a weekend, having value true if so, and false if it is a workday. The *flag_holiday* has also been added, providing information on whether the current date represents a holiday or not. The value of the *flag_holiday* feature, for a respective measurement, is given depending on the importance of the holiday, as we can see previously presented in Table 1. Also, a *flag_event* was included, having value 1 when the day of the measurement corresponds to one of the date intervals previously mentioned in Sect. 3.5.

Two columns identifying vacation periods were also added, the first being *flag_vacationperiod* representing the festive period when several people take their holidays. This flag as value 1 for every measurement made between 22 December 2018 to 1 January 2019 and 0 for every other case. The second *flag_vacationschool* represents the vacation period from school, that lasts approximately between 14 December 2018 to 31 January 2019, where once again, all measurements made inside this time period have value 1, and the rest has a value of 0.

The latest flags added represented 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, mist, haze or smoke. The *flag_rain* was also added giving information if during the measurement it was raining having value 1, and 0 if not. The same happens with the *flag_storm* representing whether a thunderstorm was occurring during the measurement.

To conclude, we have total of 2950 rows for the complete dataset, composed by 23 columns, more specifically the following ones: *temp*, *temp_min. temp_max*, *flag_rain*, *flag_fog*, *flag_storm*, *year*, *month*, *day*, *hour*, *flag_weekend*, *flag_event*, *flag_holiday*, *flag_vacationperiod*, *flag_vacationschool*, *road_tunel_marques*, *road_mouzinho*, *road_duque*, *road_castilho*, *road_braamcamp*, *road_herculano*, *road_praca_marques*, *tax_occupation.*

4.2 Algorithm Testing

In this section we started to perform algorithms tests to predict the total occupation of the parking lot. The algorithms chosen to test this were Neural Networks

(NN) like in [16], has it is a good solution for a time series prediction like our problem of parking occupancy. Distributed Random Forest (DRF) were also used, as those are a forest of classification or regression trees, rather than a single classification or regression tree, as in [18], where Regression Trees had better results and less computationally needs comparing to SVR and NN. At last, Gradient Boosting Machine (GBM), that corresponds to a sequentially regression trees on all the features of the dataset in a fully distributed way, being a good option as those gave the best results in [6]. Those algorithms were applied to the complete dataset, where the datasets where divided in 70% to train and 30% to test and to validate the results obtained we used a 5-fold cross validation, has it helps to prevent the over-fitting [18]. To test the impact of each feature type, we decided to test each of these feature in the complete dataset (Events, Traffic and Weather) with the park occupancy data, as well as the park occupancy data alone. All these algorithms were implemented with the help of the python library H2O [5]. We then tested the algorithms with the complete dataset, resulting in the accuracy metric and the execution time metric values shown in Table 3.

Table 3. Accuracy and mean execution time results.

Dataset	GBM	Execution time	DRF	Execution time	NN	Execution time
Completed Dataset	72%	10 s	71%	13 s	67%	25 s
Park + Events Dataset	76%	5 s	76%	9 s	60%	23 s
Park + Weather Dataset	74%	5 s	73%	10 s	61%	21 s
Park + Traffic Dataset	73%	6 s	70%	9 s	65%	24 s
Park Dataset	69%	3 s	71%	8 s	53%	24 s

The results obtained in Table 3, show that overall the GBM model had the best results in terms of accuracy, with a maximum 76% accuracy result with the park plus events dataset, where the RDF model showed an equal value of 76% and the NN model 60%, respectively. We were also able to verify that for all the other datasets, we obtained very uniform results, where GBM reached values no lower than 69% and with a maximum of 76%. For the DRF model we obtained values as low as 70% when using the park plus weather dataset and as high as 76% using the park plus the events dataset. In the case of the NN model the complete dataset have shown the best result with a total of 67% accuracy and, for the case of the lowest accuracy, the park dataset showed again the worst accuracy results, with only 53%. One of the possible reasons for the accuracy values of the NN 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, with only 2950 rows.

Some reasons for the results obtained and the features not giving much more information in terms of predicting the park occupancy, could be due to three reasons. The first being the parking lot categorised as an office category park, where people will 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 is that the user with a covenants will park 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.

Being this a solution to be implemented in the application the execution time is important for the application efficiency and speed. When analyzing this metric, we see that GBM overall reveals the lowest executions times with a maximum of 10 s to run while using the complete dataset, being this the largest dataset. In comparison, with the complete dataset the RDF needs 13 s and the NN needs 25 s. The GBM model also had the lowest execution time with only 3 s needed and an average of 6 s for all of the cases. Like the GBM model, the DRF model also show low mean execution times with an average of 10 s and for the case of the NN model we see the biggest execution times with an average execution time for all datasets of 22 s.

Taking those conclusions into consideration, we choose the GBM model with the complete dataset as the model to be exported and used in the android application to provide the parking occupancy levels for a certain time interval. Although this is not the model with the best values of accuracy and execution time, it does not depend 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

Table 4. Confusion matrix for the GBM model built with the complete dataset.

0	10	20	30	40	50	60	70	80	90	100	Error	Rate
85	51	0	1	0	0	0	0	0	0	0	0.3795	52/137
9	1497	13	1	0	0	1	0	0	0	0	0.0158	24/1521
0	55	200	6	2	2	1	0	0	0	0	0.2481	66/266
0	15	12	111	10	2	0	0	0	0	0	0.2600	39/150
0	5	6	13	77	3	2	1	0	0	0	0.2804	30/107
0	2	2	0	3	45	1	1	2	2	0	0.2241	13/58
0	5	3	0	0	4	118	1	2	2	0	0.1259	17/135
0	2	0	0	0	1	6	55	8	3	1	0.2763	21/76
0	0	0	0	0	0	4	3	132	18	1	0.1646	26/158
0	0	0	0	0	0	1	2	18	263	3	0.0836	24/287
0	0	0	0	0	0	0	0	1	16	38	0.3091	17/55
94	1632	236	132	92	57	134	63	163	304	43	0.1115	329/2950

compromised and the difference between the best accuracy and the accuracy of the complete dataset is only 4%.

After that we decided to analyze the confusion matrix of the GBM Model built with the complete dataset, as the results can be see in Table 4.

By analyzing Table 4, we can see that the dataset is unbalanced because it has almost half of the records at the expected value of 10%. To try and avoid this problem of unbalanced, we aggregated certain values with fewer counted records to try to balance the categories with more data. In this case, we then divided the occupation of the park into categories of 0–10%, 20–30%, 40%–50%, 60%–70% and 80%–90% and 100%, thus balancing the number of each category and, in turn, making it easier to predict new values within those categories. w in our case we implemented the above categories. After this treatment, we obtained an increase on accuracy levels to a total of 79%, an increase of 7%, and an execution time of 6 s, having an increase in performance of 4 s. The new results for the confusion matrix can be seen in Table 5.

Table 5. Confusion matrix for the GBM model built with occupancy categorization.

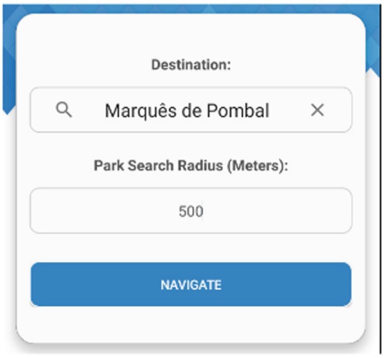
0%–10%	20%–30%	40%–50%	60%–70%	80%–90%	100%	Error	Rate
1212	51	1	1	0	0	0.0419	53/1265
84	615	15	1	0	0	0.1400	100/715
1	32	187	6	1	0	0.1762	40/227
3	12	5	169	12	5	0.1996	37/206
0	1	0	8	299	29	0.1101	37/336
0	0	0	0	28	173	0.1393	28/201
1300	711	208	184	340	207	0,1	295/2950

As we can see by evaluating the results in Table 5, we have a better distribution of the results than in Table 4, so thus increasing the rate of correct forecasts. To conclude the model being used in the rest of study it is built using the GBM model with the help of the complete dataset, to predict one of previously presented six categories. These models are built using historical information with an interval of one hour from the parking lot occupancy, weather data, traffic data and events data.

5 Results Evaluation

In this section we focused on the consolidation of the proposed system and test its efficiency and functionality. For this assessment we considered the example where a user is trying to find an available place at a peak hour in Lisbon.

For this example, the user started at ISCTE (Instituto Superior de Ciências do Trabalho e da Empresa) in Lisbon, looking for a place to leave the car in the area of Marquês de Pombal with a radius of 500 m of search between the destination and a parking lot, where all the parks outside of this range will be discarded. If there is not a single option inside the range defined by the user, the application will suggest the closest parking lot and a pop up will appear asking the user if the application should show the route to that park. This search will be performed at the 9:00 am to simulate the peak hour on the area of Marquês de Pombal.



Destination:

Marquês de Pombal

Park Search Radius (Meters):

500

Fig. 2. Navigate option with Marquês de Pombal as the destination.

As we can see in the Fig. 2, the user inserted Marquês de Pombal on the first input box and 500 m on the second input box. After this, the user clicked on navigate and the decision algorithm would run to calculate the weight of each option. This algorithm takes into consideration four heuristics to find the best suitable parking lot. First it takes into account the occupancy of the parking lot at the time of arrival, secondly the duration it takes from the current location of the user to the park, next the distance from the parking lot to the destination provided by the user, in this case Marquês de Pombal and lastly, the price per hour for having the vehicle parked in the parking lot. By taking this information the algorithm was executed and provided the result of each option, while also giving the value for the weight feature. The option with the biggest value on the weight feature will be chosen and the route to that parking lot will be created.

After executing the navigate option, the decision algorithm will run and decide which of the parking lots is the best option, in this case we only have

Table 6. Decision algorithm heuristics and weight outcome.

Parking lot	Route duration (s)	Distance to the parking lot (m)	Distance from Destination to parking lot (m)	Parking lot occupancy (Percentage)	Price per hour (€)	Weight
Park 1	939	3029	348	40%–50%	2.15	184.85

one parking lot added, so the output of the decision algorithm can be seen in Table 6.

By analyzing Table 4, we can see that the occupancy at the time of arrival is between 40%–50%, being this the most important heuristic to take into consideration by the decision algorithm. The duration from the current location to the destination is also an important heuristic to take into consideration and in this case the user was at 939 s away (15 min and 39 s). For the third heuristic to take into consideration we have the distance from the destination and the parking lot chosen, in this case the parking lot is at 348 m away from the destination provided by the user. At last, the algorithm takes into consideration the price per hour it costs the user to have their vehicle in the parking lot and for the parking lot in consideration it is 2.15€. By providing all these features to the decision algorithm it will produce an output (weight), which will be used for the decision of the best parking lot, as the best option will have the highest weight value, and the route to that vehicle park will be created, as we can see in Fig. 3.

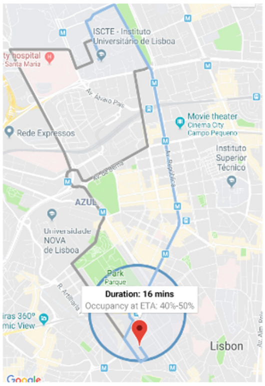


Fig. 3. The route to the Park 1 chosen from the decision algorithm.

6 Conclusions

This work provides a solution to real-time guidance to parking places in a general solution to the problem that is increasingly real in large cities which is to find a free place to park the vehicle. A solution like the one shown has the potential to reduce traffic within cities and in turn allow users to better manage their time and route. In this study we consider three types of data to help predict the occupation of a park categorized as an office park, namely events, traffic and weather in the vicinity of the park, concluding that the data types with the greatest impact are the events, mainly in the rotation type users. We were also able to conclude that the traffic data show a high correlation with the occupation situation of the park. This park contains the option of covenants, which does not suffer much disturbance, since the user has already paid a fee to be able to guarantee the car inside the park over a given period, forcing the user in a certain way to leave the car in the same parking lot. Good accuracy values were obtained, however, the fact that the data are short and that the month of December has patterns of parking not very common, due to the fact that it is a festive month and with many holidays, can condition the forecast values. As a future work it would be interesting to add more car parks to understand how the decision algorithm behaves in several cases and to understand the impact of the same types of data on those same car parks.

References

1. Bock, F., Sester, M.: Improving parking availability maps using information from nearby roads. Transp. Res. Procedia **19**(June), 207–214 (2016). https://doi.org/10.1016/j.trpro.2016.12.081
2. Caliskan, M., Barthels, A., Scheuermann, B., Mauve, M.: Predicting parking lot occupancy in vehicular ad hoc networks. In: 2007 IEEE 65th Vehicular Technology Conference - VTC 2007-Spring, pp. 277–281 (2007). https://doi.org/10.1109/VETECS.2007.69. http://ieeexplore.ieee.org/document/4212497/
3. Chen, B., Pinelli, F., Sinn, M., Botea, A., Calabrese, F.: Uncertainty in urban mobility: predicting waiting times for shared bicycles and parking lots. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC (ITSC), pp. 53–58 (2013). https://doi.org/10.1109/ITSC.2013.6728210
4. Giuffrè, T., Siniscalchi, S.M., Tesoriere, G.: A novel architecture of parking management for smart cities. Procedia - Soc. Behav. Sci. **53**, 16–28 (2012). https://doi.org/10.1016/j.sbspro.2012.09.856. https://linkinghub.elsevier.com/retrieve/pii/S1877042812043182
5. H2O: H2o documentation (2019). http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html. Accessed 8 Apr 2019
6. Ionita, A., Pomp, A., Cochez, M., Meisen, T., Decker, S.: Where to park?: predicting free parking spots in unmonitored city areas. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, pp. 22:1–22:12 (2018). https://doi.org/10.1145/3227609.3227648
7. Klappenecker, A., Lee, H., Welch, J.L.: Finding available parking spaces made easy. Ad Hoc Netw. **12**(1), 243–249 (2014). https://doi.org/10.1016/j.adhoc.2012.03.002
8. Lijbers, J.: Predicting parking lot occupancy using prediction instrument development for complex domains (2016)
9. OpenWeatherData: Openweatherdata history bulk (2019). https://openweathermap.org/history-bulk. Accessed 23 Apr 2019
10. Pflügler, C., Köhn, T., Schreieck, M., Wiesche, M., Krcmar, H.: Predicting the availability of parking spaces with publicly available data. Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn, pp. 361–374 (2016). http://cs.emis.de/LNI/Proceedings/Proceedings259/361.pdf
11. Pullola, S., Atrey, P.K., Saddik, A.E.: Towards an intelligent GPS-based vehicle navigation system for finding street parking lots. In: ICSPC 2007 Proceedings - 2007 IEEE International Conference on Signal Processing and Communications (November), pp. 1251–1254 (2007). https://doi.org/10.1109/ICSPC.2007.4728553

12. Richter, F., Martino, S.D., Mattfeld, D.C.: Temporal and spatial clustering for a parking prediction service. In: Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI 2014-December, pp. 278–282 (2014). https://doi.org/10.1109/ICTAI.2014.49
13. Rong, Y., Xu, Z., Yan, R., Ma, X.: Du-parking: spatio-temporal big data tells you realtime parking availability. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 646–654 (2018). https://doi.org/10.1145/3219819.3219876
14. Shin, J.H., Jun, H.B.: A study on smart parking guidance algorithm. Transp. Res. Part C: Emerg. Technol. **44**, 299–317 (2014). https://doi.org/10.1016/j.trc.2014.04.010
15. Tilahun, S.L., Di Marzo Serugendo, G.: Cooperative multiagent system for parking availability prediction based on time varying dynamic Markov chains. J. Adv. Transp. **2017** (2017). https://doi.org/10.1155/2017/1760842
16. Vlahogianni, E.I., Kepaptsoglou, K., Tsetsos, V., Karlaftis, M.G.: A real-time parking prediction system for smart cities. J. Intell. Transp. Syst.: Technol. Plan. Oper. **20**(2), 192–204 (2016). https://doi.org/10.1080/15472450.2015.1037955
17. Xiao, J., Lou, Y., Frisby, J.: How likely am I to find parking? - a practical model-based framework for predicting parking availability. Transp. Res. Part B: Methodol. **112**, 19–39 (2018). https://doi.org/10.1016/j.trb.2018.04.001
18. Zheng, Y., Rajasegarar, S., Leckie, C.: Parking availability prediction for sensor-enabled car parks in smart cities. In: 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 7–9 April (2015)