MEF UNIVERSITY

PREDICTION OF HOTEL RESERVATION CANCELLATIONS

Capstone Project

Oğuz Kirazdiken

ISTANBUL, 2020

MEF UNIVERSITY

PREDICTION OF HOTEL RESERVATION CANCELLATIONS

Capstone Project

Oğuz Kirazdiken

Advisor: Asst. Prof. Duygu TAŞ KÜTEN

ISTANBUL, 2020

MEF UNIVERSITY

Name of the project: Prediction of Hotel Reservation Cancellations
Name/Last Name of the Student: Oğuz Kirazdiken
Date of Thesis Defense: /09/2020

I hereby state that the graduation project prepared by Oğuz Kirazdiken has been completed under my supervision. I accept this work as a "Graduation Project".

/09/2020
Asst. Prof. Duygu TAŞ KÜTEN

I hereby state that I have examined this graduation project by Oğuz Kirazdiken which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

/09/2020

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Duygu TAŞ KÜTEN	……..
2. Prof. Dr. Özgür ÖZLÜK	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name Date Signature

EXECUTIVE SUMMARY

PREDICTION OF HOTEL RESERVATION CANCELLATIONS

Oğuz Kirazdiken

Advisor: Asst. Prof. Duygu TAŞ KÜTEN

SEPTEMBER, 2020, 24 pages

This study aims to create a model to predict the reservation requests that may be potentially cancelled using the information related to reservation requests. For this purpose, the reservation request gathered by hotels in Portugal for a period of two years are examined. First, the features that may contain important information for reservation cancellations are specified. These features are then used in the training and testing stages of the classification models employing several preprocessing phases to explain the reasons of reservation cancellations. Machine learning algorithms such as random forest, decision trees, logistic regression, gradient boosting classifier and extra gradient boosting classifier are trained and tested on the above-mentioned real data and the results are discussed in detail.

Key Words: Hotel Reservation Cancellations, Classification, Random Forest Classification, XGBoost Classification.

ÖZET

OTEL REZERVASYON İPTALLERİNİN TAHMİNLENMESİ

Oğuz Kirazdiken

Tez Danışmanı: Dr. Öğretim Üyesi Duygu TAŞ KÜTEN

EYLÜL, 2020, 24 Sayfa

Bu çalışmada, bir otele gelen rezervasyon taleplerinden hangilerinin iptal edileceğini rezervasyon detayındaki çeşitli bilgilere bakarak tahminleyebilecek bir model oluşturmak amaçlanmıştır. Bu amaç doğrultusunda Portekiz'de bulanan otellerden iki yıl boyunca toplanan rezervasyon talepleri incelenmiştir. Öncelikle, rezervasyon iptallerini açıklayabilmek adına önemli bilgiler içeren rezervasyon detayları belirlenmiştir. Daha sonra bu özellikler, rezervasyon iptallerinin nedenlerini açıklamak için birkaç ön işleme aşaması kullanan sınıflandırma modellerinin eğitim ve test aşamalarında kullanılmıştır. Rastgele orman, karar ağacı, lojistik regresyon, sınıf takviyeli sınıflandırıcı ve extra sınıf takviyeli sınıflandırıcı gibi makine öğrenmesi algoritmalarında r oluşturulmuş ve sonuçları karşılaştırılmıştır.

Anahtar Kelimeler: Otel rezervasyon iptalleri, Sınıflandırma, Rastgele Orman Sınıflandırma, Extra Sınıf Takviyeli Sınıflandırma.

TABLE OF CONTENTS

1. INTRODUCTION

In service industries, especially in hospitality sector, the difference between booking demands and realized accommodations may lead misuse of resources and cause companies a considerable amount of operational costs. This issue increases the importance of correct use of resources and revenue management for companies operating within the sector. Revenue management is the allocation of on-hand funds at the right time with correct amount and price (Kimes and Wirtz, 2003). It plays an important role in forecasting demand, since it allows to calculate multi-component demand problems in a very comprehensive way (Talluri and Ryzin, 2009). In the service sector, calculating real demand for a specific time period is a complicated problem affected by several variables. In order to predict the net amount of accommodation that will take place within a time frame, it is necessary to estimate which booking requests may be cancelled (Lemke, Riedel and Gabrys, 2013).

Reservation cancellations may correspond to huge percentages for some businesses with respect to the total number of reservations and that ratio may lead hotels to enforce strict cancellation rules or narrow their overbooking strategies (Chen, 2016). The policies applied in both cases may be a factor that affecting the customers' preferences for current or future accommodation. Not well-established booking strategies may lead to having unsatisfied customers who may never book again and crucial revenue losses for enterprises (Mehrotra and Ruttley, 2006). Although some of the reservation cancellations are due to natural causes such as bad weather, activity or travel cancellations, most of them are due to customers seeking for a better deal (Chen, Schwartz and Vargas, 2011). Firms that have understood the true causes of their cancellations can offer more dynamic cancellation options according to customer selections which strength their net demand forecasts and allow them to obtain a better financial status (António, Almeida and Nunes, 2019). A predictive model on cancellations can be quite helpful to forecast the real demand for a specific period. To provide an accurate forecast of demand in the hospitality sector, this project aims to create a model predicting whether the accommodation will take place or not according to the information obtained from customers' booking requests.

1.1. Literature Survey

There are studies on the forecasting of reservation cancellations as mentioned in this section. In fact, studies similar to the models and predictions used in this study have been carried out on plane tickets and restaurant visits as well. However, according to Benítez Aurioles (2018), there is only a little research done on this topic, especially on hotel industry.

The choice of the channel to be booked is not arbitrary and independent of the realization of the booking plan, but on the contrary it depends on individual and technical circumstances. (Dolnicar and Laesser, 2007). For example, travel budget, additional personnel requirements, and lead time are an example of these circumstances (Law, Leung, and Wong, 2004). As per Cheyne, Downes, and Legg (2006) spontaneously made, short-term and affordable mini trips are mostly booked through online channels. On the other hand, people who plan long-term vacation prefer travel agencies since there are more than one variable to consider constituting a high degree of complexity of booking. On the different study of Chew and Jahari (2014), it is found that when there is a long time between the booking and the hotel registration date, circumstances that leading to hotel cancellations such as dinner plans planned at the last minute, natural disasters, unpredictable internal and external events that are not fully considered on the date of booking (Fesenmaier and Jeng, 2000) etc. are more likely to happen. All the above-mentioned authors arrive at similar conclusion which is they all have started their work by evaluating and sampling various features while modelling the cancellation behavior of hotel guests.

In the literature, most studies consider the above-mentioned problem as a regression problem which need to exhibit the total cancellation rate in a specific time period according to the general values of a business. Seminal studies such as Morales and Wang (2010), mention the difficulty of calculating the cancellation probability of each reservation request, and also cited that even if it is calculated, it is not possible to reach the net accommodation numbers with sufficient accuracy. However, as Ivanov (2014) states that estimating the net amount of accommodation that will take place within a certain period of time is only possible with building an accurate relationship between reservation details such as customer type and market segment with cancellations. There are also studies considering the problem as a classification problem, such as Antonio, Almeida and

Nunes (2017). The classification method tries to predict the requests to be cancelled using the descriptive information obtained from the details of the reservation requests. All these studies differ in terms of handling the model. As mentioned above, the problem has been handled in two different ways as regression model or classification model. However, the common point of all studies is to use the details of the reservation information, regardless of the model type.

Apart from this, many different studies have been made on the reasons that cause the cancellations and on the estimation of cancellations in the plane ticket purchases (Garrow and Koppelman, 2004) and restaurant visits (Tse and Poon, 2016).

2. PROJECT DEFINITION

2.1. Problem Statement

It is very important to determine the amount of net accommodation that will take place within a certain time frame in the service industry. Almost all functions of companies operating in service industry make necessary organizational preparations such as budget allocation and product supply according to the number of guests. Large differences between expected and actual accommodation causes these functions to misuse their resources.

Companies that try to keep reservation cancellations at a certain level may set strict rules during the creation of the reservation requests. However, these rules may lead customers to prefer another accommodation place in today's competitive conditions in service industry. Understanding the real reasons of reservation cancellation helps companies to create much more efficient reservation rules which can prevent customer loss. Companies which estimate the net accommodation amount to be realized in a certain period accurately can use their own resources more efficiently.

2.2. Project Objectives

The purpose of this project is to determine which reservation details can explain the reservation cancellations more accurately and to create the appropriate classification model with determined inputs. For this purpose, certain preprocessing phases are applied to the reservation details which are considered important in order to predict cancellations. Accordingly, different classification models are implemented with these inputs. Models which can predict net accommodation amounts with high accuracy are fine-tuned and the results are explained in detail.

3. ABOUT THE DATA

During the project, The Hotel Booking Demand Dataset will be used for the analysis of required features due to its wide range of information (Mostipak, 2020). The dataset includes booking requests of two hotels in Portugal for a time period of two years from June 2015 to August 2017 and the information of the requests that have been canceled. Working with 31 different booking details, such as the date of the booking was made, the length of the stay, the number of adults, children, and/or babies, and the number of available parking spaces, contribute to study during the creation process of the model to get more accurate outputs. Instead of establishing a detailed model with the historical data of a single place, creating a model including more generic factors of cancellation process with more than one enterprise's data may provide more successful results (António, Almeida and Nunes, 2019).

3.1. Features

The features of dataset and their explanations are listed as follows:

- Hotel: Relevant hotel type (Resort Hotel or City Hotel).
- Is Canceled: The information label about whether the reservation is canceled or not.
- Lead Time: Number of days between the date of reservation was made and the date of stay.
- Arrival Date Month: The arrival month of stay.
- Arrival Date Week Number: The week of year when the accommodation has begun (1 – 52).
- Arrival Date Day of Month: The day of month when the accommodation has begun (1 – 31).
- Stays in Weekend Nights: The number of weekend nights that the accommodation includes.
- Stays in Week Nights: The number of weekday nights that the accommodation includes.
- Adults: The number of adults is in the reservation request.

- Children: The number of children are in the reservation request.
- Babies: The number of babies are in the reservation request.
- Meal: Type of meal that is requested by guest in reservation (BB, FB, HB or SC).
- Country: Country codes of guests with ISO representation.
- Market Segment: Information about the market segment that made the reservation. “TA” refers to “Travel Agents” and “TO” refers to “Tour Operators”.
- Distribution Channel: Information about the market segment that made the reservation.
- Is Repeated Guest: Information of whether the guest has made any accommodation before.
- Previous Cancellations: Information about how many reservations have been cancelled by the guest before.
- Previous Bookings Not Canceled: Information about how many reservations have not been cancelled by the guest before.
- Reserved Room Type: The type of the room that the guest requested.
- Assigned Room Type: The type of the room that the guest assigned by the hotel.
- Booking Changes: Number of changes made in booking details from the first entry till the check-in or cancellation.
- Deposit Type: Information about the deposit option that the guests chose in booking (No Deposit, Non-Refund, Refundable).
- Agent: Information about the travel agency that made the reservation.
- Company: Information about the company that made the reservation.
- Days in Waiting List: Number of days passed from the moment that the reservation is made till it is confirmed.
- Customer Type: Four different categories created to describe the guest types (Transient, Contract, Group, Transient-Party).
- ADR: Average daily rate.
- Required Car Parking Spaces: Required parking space specified by the guest.
- Total of Special Requests: Number of special requests specified by the guest such as twin beds or high floor.

- Reservation Status: Information of the last reservation status (Canceled, Check-Out, No-Show)
- Reservation Status Date: Date on which the last reservation status was stated.

3.2. Exploratory Data Analysis

For the success of targeted predictive model, the features of customers and accommodation places and the effects of seasonality should be analyzed in detail. First, each feature contributes to the accuracy of the model with a different level of importance. Moreover, their ability to explain the characteristic behaviors of the reservation cancellations varies according to the accommodation type (Antonio, Almeida and Nunes, 2017). Therefore, a dataset which includes several information about the booking detail contributes greatly to the accuracy of the predictive model.

In our dataset there are 119,390 reservation records for two different hotels and the total number of cancelled records are 44,224.

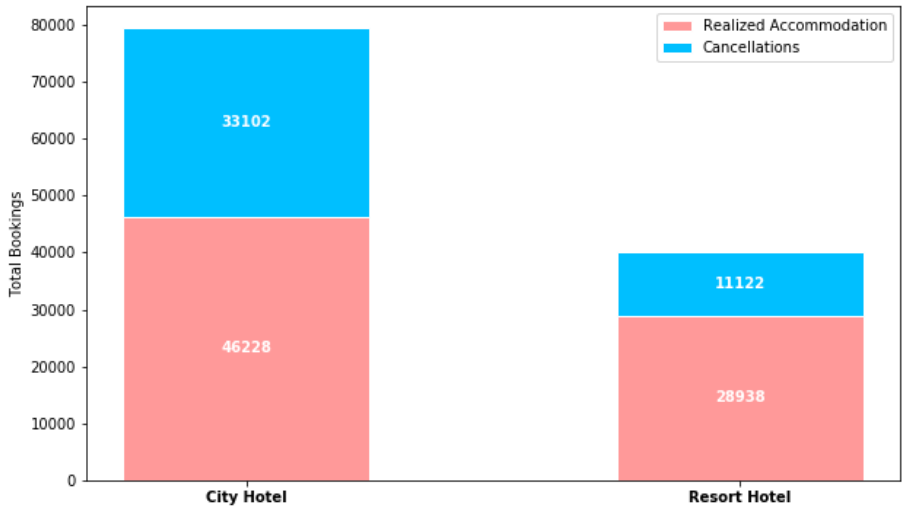

Figure 1: The number of realized accommodations and cancellations for each hotel type over the selected two-year period.

In order to understand the variables that correctly explain the customers' behaviors on the cancellations, it is important to work with a sufficient number of observations for both cancellation and show-up cases in each hotel booking data. As can be seen in Figure 1, there are enough cancelled and realized accommodation bookings in the dataset for both hotels.

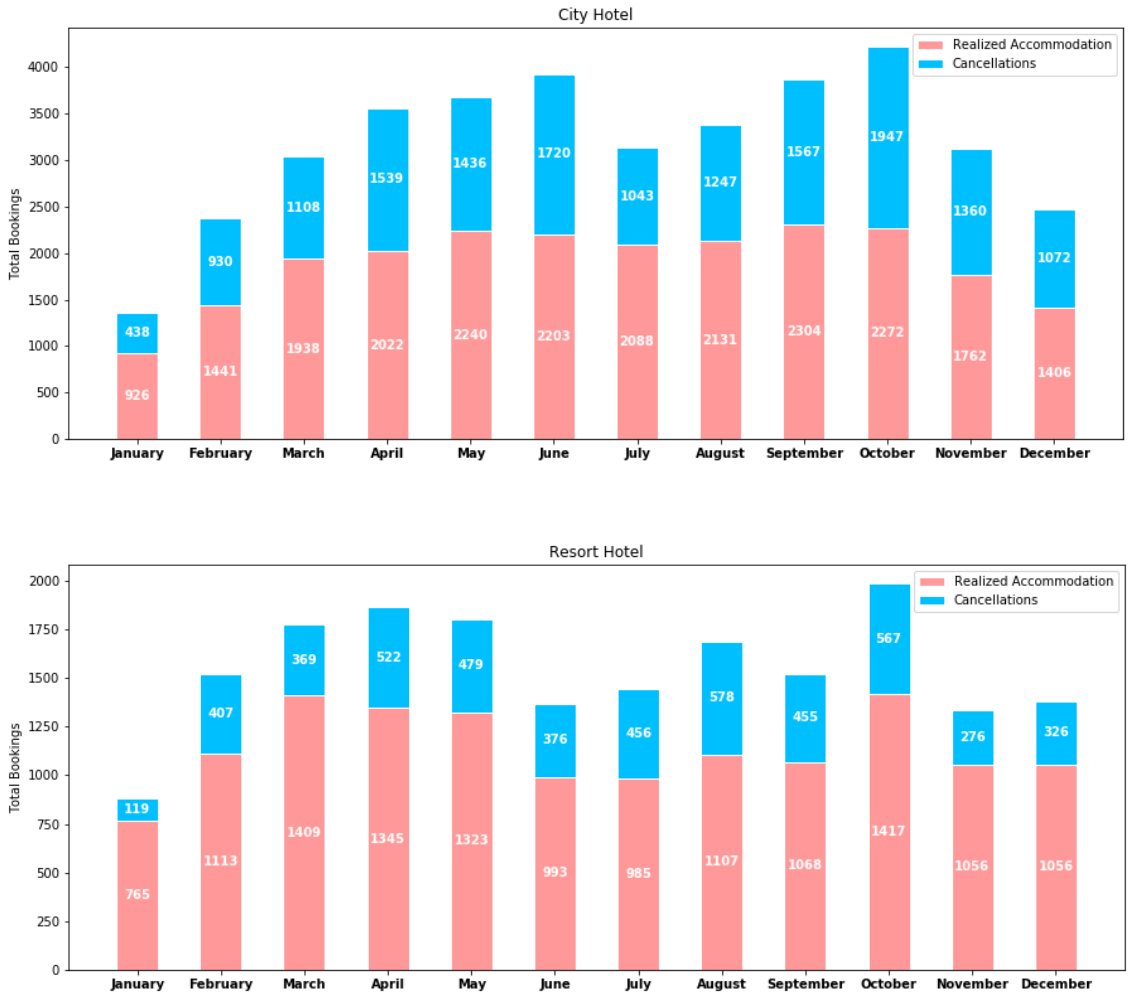

Figure 2: The number of realized and cancelled accommodations for both hotels in 2016.

To observe the monthly changes in total reservation numbers, the information of all 12 months in a year is only available for the records of year 2016 within the dataset. As seen in Figure 2, the monthly changes in the total number of reservations during the year shows similarity for both hotels. Booking requests start to increase in both hotels in February and reach their highest value in October. It is seen that the total amount of reservation requests of City Hotel is higher than the Resort Hotel for each month. Evaluating the estimates of the model according to the values shown in Figure 2 on a monthly basis and prioritizing to keep the difference of actuals and estimates in a reasonable level may increase the success of the model.

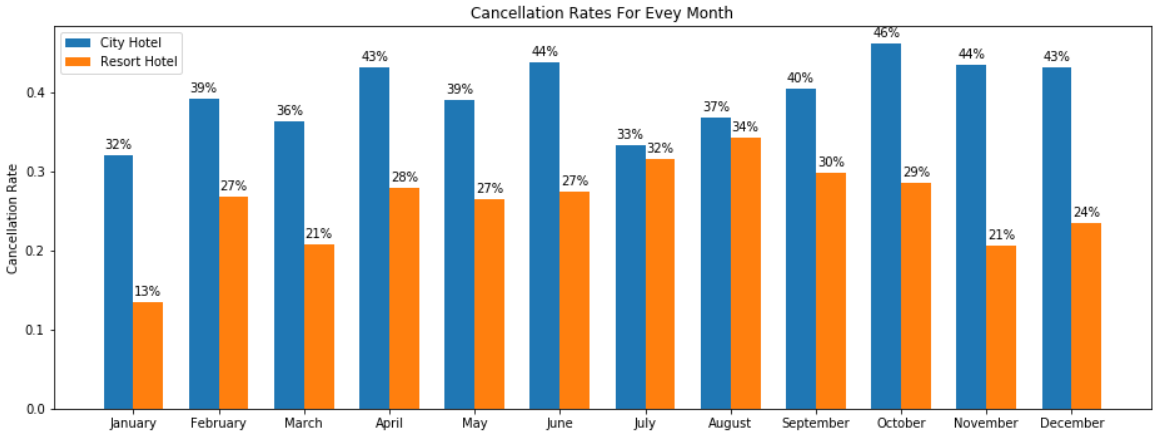

Figure 3: The change of the rate of cancelled bookings to total bookings during the year 2016 for both hotels.

Figure 3 shows the ratio of the monthly reservation cancellations to total reservations over each month. First, we observe that the percentage of reservation cancellations in City Hotel is higher than the Resort Hotel for each month. Another important observation is that while the total number of reservation requests for both hotels are increasing, the percentage of the total cancellations are also increasing for the same months. Comparing the estimates of predictive model and actual cancellation percentages in the same month may indicate the model accuracy and helps us determine whether there is a bias in the estimations of each month.

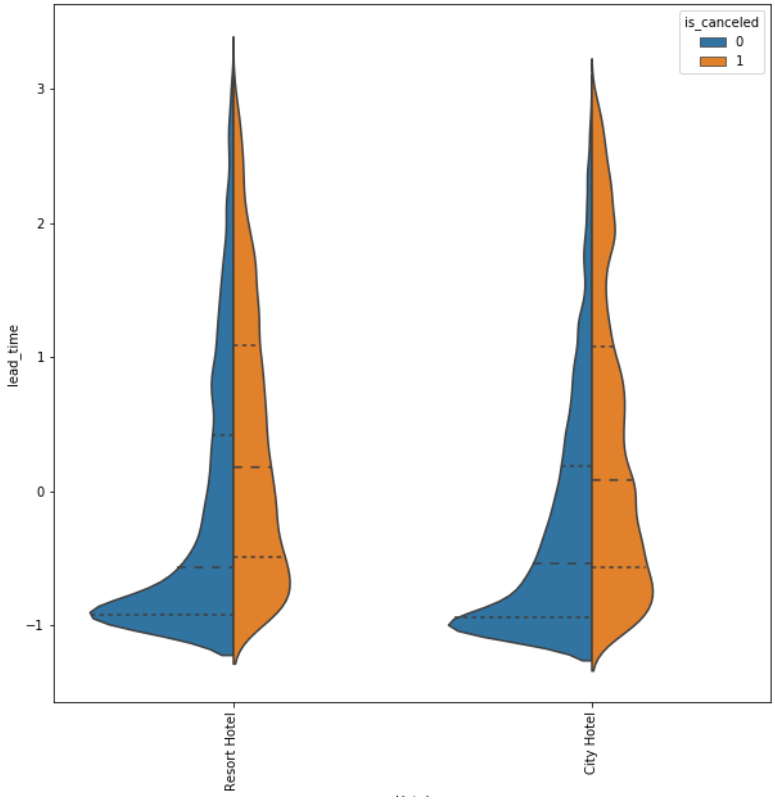

Figure 4: The distribution of standardized lead time for each label in both hotel.

Figure 4 shows the distribution of lead times which is cleared from its outliers and standardized. In total, 3005 of 11,9390 observations are evaluated as an outlier since they have greater or less values than 1.5 interquartile range distance from the first and third quartiles and not included into the graph. As it can been seen from Figure 4, there is a difference between the distribution of cancelled reservations and realized accommodations for both hotels. Such a difference indicates the power of lead time to explain the variance between cancelled and realized bookings.

Figure 5 and 6 present the change in the cancellation rates according to lead times. While calculating the cancellation rates, the number of cancelled bookings for each lead time value is divided to the total number of reservations within the same lead time value. According to the correlation graph of each hotel, it seems that there may be a relationship between the cancellation rate and lead time.

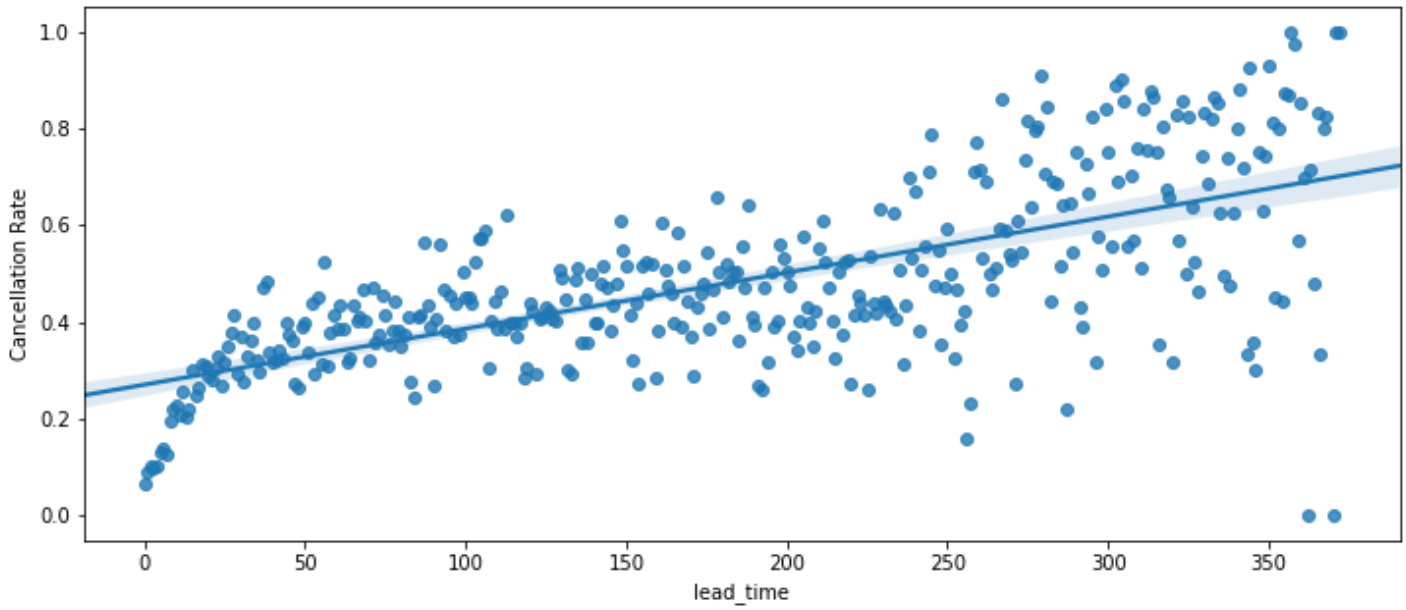

Figure 5: Distribution of cancellation rate according to lead time.

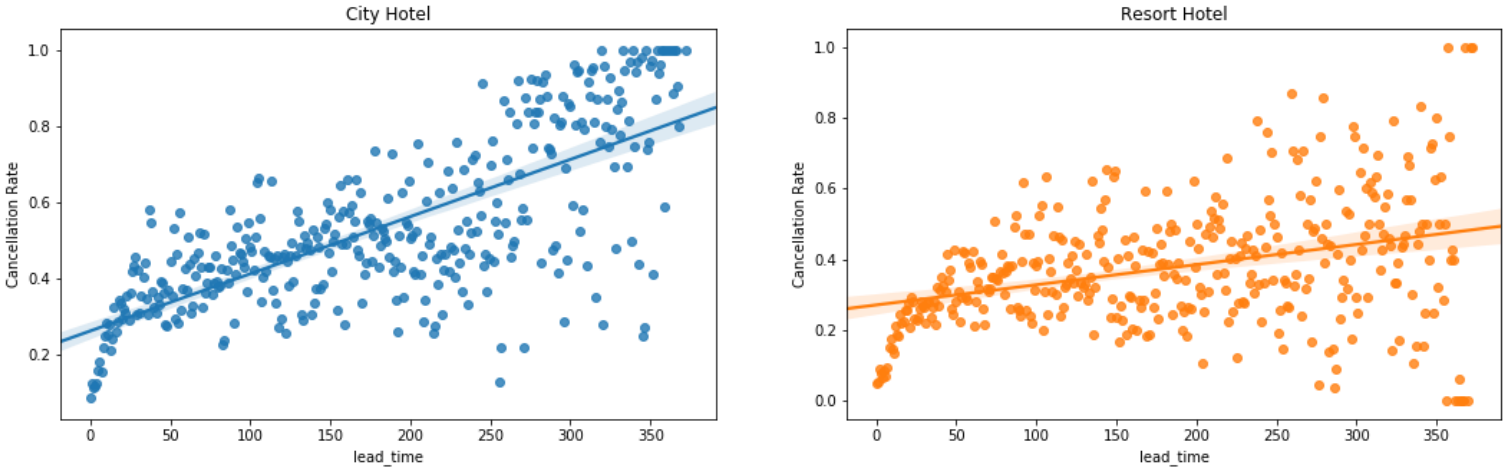

Figure 6: The distribution of cancellation rate according to lead time for each hotel

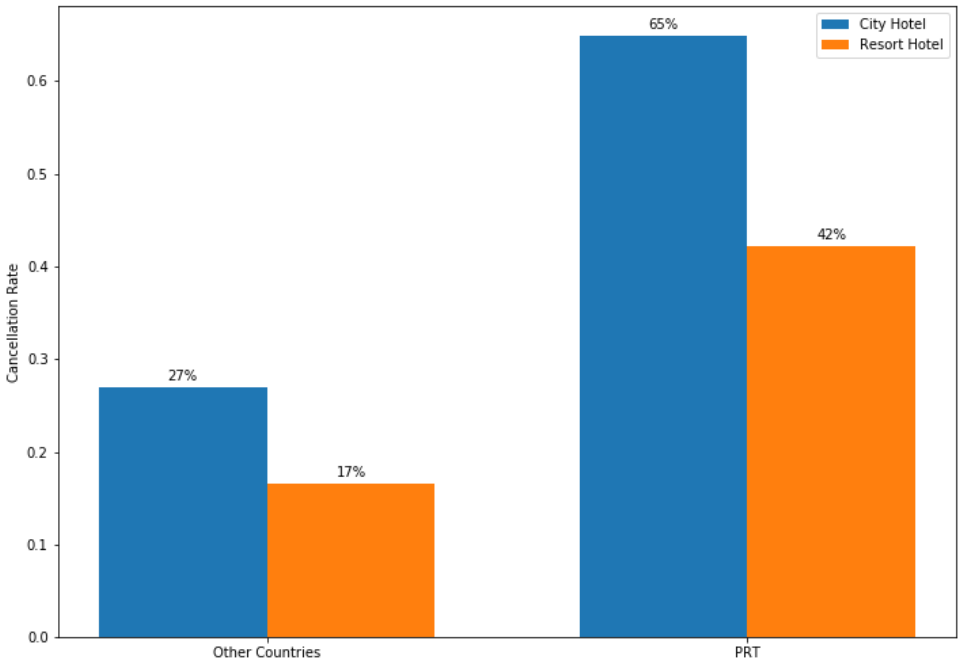

Figure 7: Cancellation rates of hotel reservations made from Portugal and other countries

Figure 7 shows the cancellation rates over all bookings made for two hotels located in Portugal from Portugal and other countries. There are 119,390 reservation requests made from 177 countries in the dataset. It is seen that while 57% of 48,590 reservation requests made from Portugal were canceled, only 24% of 70,800 reservation requests were canceled from the remaining 176 countries. Therefore, Figure 7 clearly shows that the reservation cancellation rates made from Portugal are much higher than the total reservation cancellation rates made from other countries. In this sense, information of the country from which the reservation was made may be an important data when predicting the likelihoods of reservation cancellations.

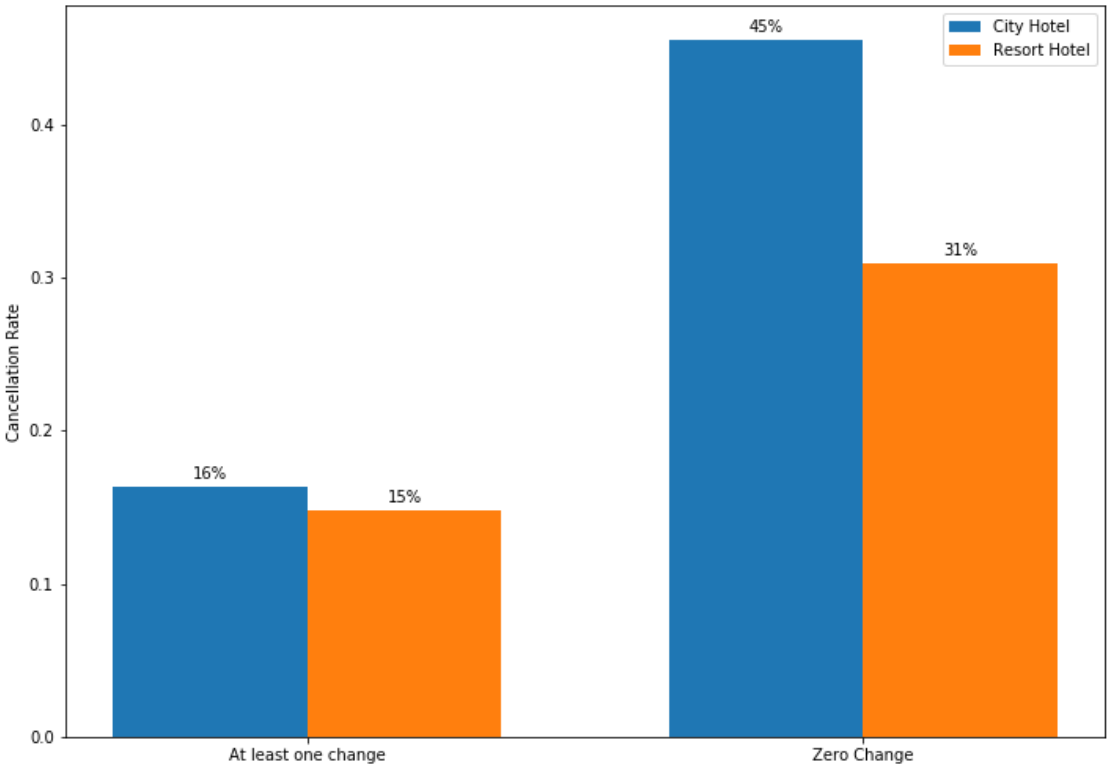

Figure 8: Cancellation rates of hotel reservations according to the information on whether the customers have updated the reservation details

Figure 8 shows the cancellation rates for two cases: (i) customers have updated the reservation details at least once after they create the requests, (ii) customers have not been changed any detail. There are 101,314 reservation requests whose details have not been changed after the date of creation and it is seen that 41% of them have been canceled. On the other hand, the remaining 18,706 requests have been changed at least once and only 16% of them has been cancelled. The significant difference in cancellation rates indicates that the likelihoods of realizing the accommodation is more for the customer who have changed reservation details at least once. In this sense, the information on whether the customers have changed the reservation details at least once may contribute the study while predicting reservation cancellations.

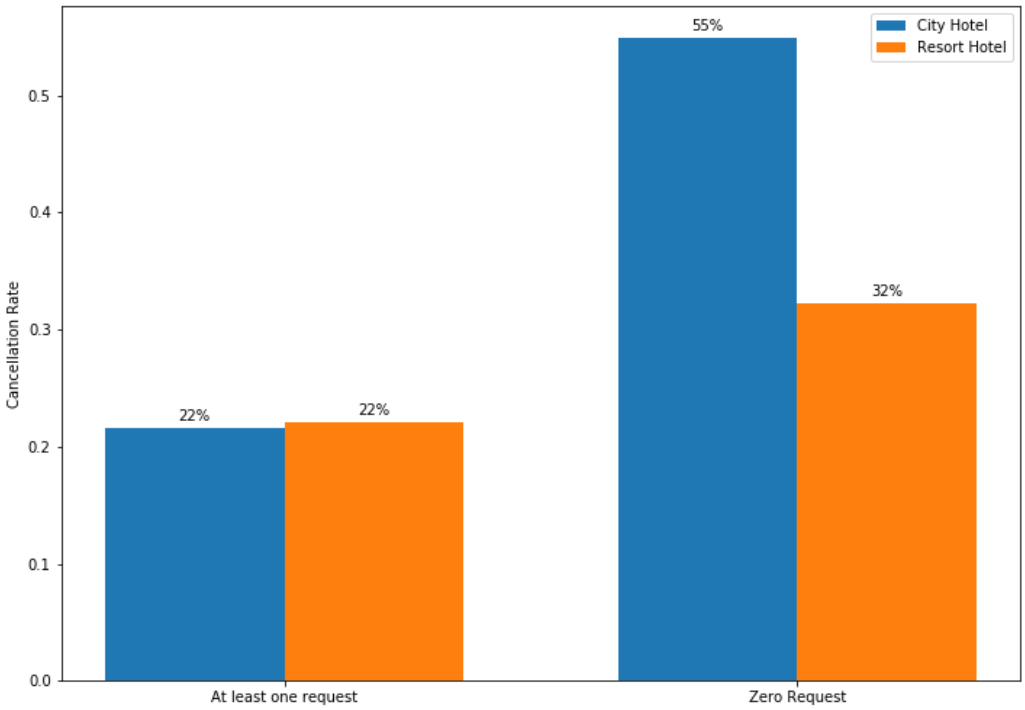

Figure 9: Cancellation rates of hotel reservations according to additional requests of customers

The information on the number of special requests that the customers have demanded is specified under the feature of “total special requests” in the dataset. Figure 9 shows the cancellation rates of bookings with at least one additional request and those with no additional requests for both hotels. There are 70,318 reservation requests in total with no additional requests in the dataset and it is seen that 48% of these requests are resulted with cancellation. However, only 22% of 49,072 bookings which has at least one additional request has been cancelled. As shown in Figure 9, bookings with at least one additional request are less likely to result with cancellation according to the evaluations made separately for each hotel. In this sense, the information on whether customers demand additional request may be beneficial while predicting reservation cancellations.

4. METHODOLOGY

The general approach in this study is to understand the details which explain the cancellations in the Hotel Booking Demand dataset, to add new features into the dataset in order to predict the cancellations more accurately and to create classification models by adapting these details. After the preprocessing studies are carried out, a number of classification models are created and those providing results with high accuracy are included in the hyper parameter tuning processes.

4.1. Data Preprocessing

In the study, “lead time” and “adr” features in the dataset are first cleaned by removing the outlier observations to prevent the model being affected by these outlier values. Therefore, 6682 observations are removed from the dataset since they are out of 1.5 interquartile range (IQR) distance from the first and third quartile values.

As mentioned in exploratory analysis (Figures 7, 8 and 9), the rate of encountering cancellations is significantly high for specific values of “country”, “booking changes” and “total of special requests” features. “Country” contains 177 different categorical values and should be included in the model after one-hot-encoding process. However, this process requires adding 177 columns into the input data of the model, inherently leading to longer computation times. Instead, a binary representation that separates the “country” from those from Portugal and other countries is beneficial for computer processing speed and does not cause any information loss. Therefore, binary representation for the “country” feature is used. Then, binary representation is also used for “booking changes” and “total of special requests” features. Next, their real values are implemented in the model and the performances of two models (with binary values and with real values) are compared in terms of model accuracy. Due to the better performance of the model employing binary values, these values are applied on the three features to increase the explanatory nature and processing speed of the model without causing any information loss.

Moreover, the number of adults, children and babies that will realize the accommodation is stated in the reservation details. However, the information that the people who will be staying are a family or not may affect the cancellation of the reservation. So, “is family” feature is added into the dataset in order to identify the

bookings of adults who have made a reservation request with at least one child or a baby. Then, the “adults”, “children” and “babies” features are removed since the newly added “is family” feature is highly correlated with these three features.

Furthermore, features with high correlations should not be included into the model together to create a suitable classification model and prevent multicollinearity. Table 1 shows two pairs of features which have the highest correlation with each other. So, each feature shown in the Table 1 is applied in the classification model one by one. As a result of this elimination process, “distribution channel” and “reserved room type” features are removed from the input dataset since it is found that the “market segment” and “assigned room type” features provide more contribution to the model in terms of accuracy.

Features		Correlation
reserved_room_type	assigned_room_type	81%
market_segment	distribution_channel	77%

Table 1: The two highest correlation pairs

The “reservation status”, “reservation status date”, “agent” and “company” features are removed from the input dataset of the model because of the following reasons: i) “agent” feature dramatically increases the computation time since it has 335 unique categorical values and processing one hot encoding does not increase the model accuracy, ii) “company” feature has lots of missing values (112,593 blank rows out of 119,390), iii) “reservation status” feature has exactly the same values as the “is cancelled” feature and iv) “reservation status date” feature is a time stamp. Overall, all these four features do not have an explanatory power to identify reservation cancellations and may create an extra computer processing load when included in the model input data.

4.2. Standardization and One-Hot-Encoding

23 features are left after the data preprocessing phase. The remaining features and their variable types are shown in Table 2. Although “arrival date year”, “arrival date week number” and “arrival date day of month” features have numeric values, they are considered as categorical variables. To only indicate the week of the year and the day of the month at which the accommodation is realized.

Feature	Type
is_family	Binary
from_portugal	Binary
is_repeated_guest	Binary
booking_changes	Binary
total_of_special_requests	Binary
hotel	Categorical
arrival_date_year	Categorical
arrival_date_month	Categorical
arrival_date_week_number	Categorical
arrival_date_day_of_month	Categorical
meal	Categorical
market_segment	Categorical
assigned_room_type	Categorical
deposit_type	Categorical
customer_type	Categorical
lead_time	Numeric
stays_in_weekend_nights	Numeric
stays_in_week_nights	Numeric
previous_cancellations	Numeric
previous_bookings_not_canceled	Numeric
days_in_waiting_list	Numeric
adr	Numeric
required_car_parking_spaces	Numeric

Table 2: Features and their variable types

To clearly build the relationship between numeric values in classification models, a standardization study is required. For this reason, the distances between the numeric feature values and their mean values are found for each observation and then, divided by their standard deviations. These standardized values of numeric features are used as model inputs rather than their old values. Then, one hot encoding study is performed for the categorical features and a new binary column for each categorical value is provided in the input dataset.

5. RESULTS

70% of reservation requests of the dataset is allocated for model training and the remaining 30% is used for model testing. Prediction models are created by using random forest classifier, decision tree classifier, logistic regression, extra gradient boosting classifier, multilayer perceptron classifier, adaptive boosting classifier and gradient boosting classifier with keeping the same 70% of observations for model training and same 30% of observations for model testing.

Table 3 shows the results of the train and test set accuracy scores of the models, the average of five different cross validation scores for the train and test sets and the area under curve (AUC) values of the receiver operator characteristic (ROC) graphs for test set predictions of each model. While creating the models, default parameters are used.

	Trainset Accuracy	**Avg. CV Score of Trainset**	**Testset Accuracy**	**Avg. CV Score of Testset**	**AUC for Testset**
RandomForestClassifier	98,76%	87,01%	87,59%	85,20%	94,00%
DecisionTreeClassifier	99,58%	84,25%	85,05%	82,43%	84,00%
LogisticRegression	82,23%	82,05%	82,25%	82,01%	90,00%
XGBClassifier	88,35%	86,73%	86,70%	86,08%	94,00%
MLPClassifier	95,52%	85,24%	85,45%	83,72%	93,00%
AdaBoostClassifier	82,38%	82,22%	82,53%	82,55%	90,00%
GradientBoostingClassifier	84,33%	84,19%	84,45%	84,31%	92,00%

Table 3: Results of the models for training and testing.

In the model created with random forest classifier, the train set is predicted with 98.76% accuracy. However, when the same train set is divided into five equal parts in the cross-validation process and put into a separate model, an average accuracy of 87.01% is estimated. As a result, it is not possible to mention any overfitting because of the difference between the model's train and the test set accuracies. Results presented in Table 3 shows that the average cross validation scores of random forest classifier and extra gradient boosting classifier provide the highest values for the test and train set. In addition, the AUC values of these two models are higher than the other models.

In addition, hyperparameter tuning process is carried out for random forest classifier and it has been observed in the first model implementation that the extra gradient boosting classifier gives the best results.

5.1.1 Hyperparameter Tuning

In the study performed on the GridSearchCV algorithm, a number of parameters are tested for the random forest classifier and for the extra gradient boosting classifier models. For the random forest classifier model, the dataset is fitted into the model with different parameter combinations for 1728 times and the best parameters for the highest accuracy is reported in Table 4 as a result of GridSearchCV study.

GridSearchCV Output	Parameters After Manual Iterations
'bootstrap': False	'bootstrap': False
'criterion': 'entropy'	'criterion': 'entropy'
'max_depth': 50	'max_depth': 40
'min_samples_leaf': 1	'min_samples_leaf': 1
'min_samples_split': 5	'min_samples_split': 5
'n_estimators': 700	'n_estimators': 1000

Table 4: Output of GridSearchCV for random forest classifier and parameters obtained after manual iterations.

Similarly, GridSearchCV algorithm is also used for the hyperparameter tuning study of the extra gradient boosting classifier model. There are 864 different combinations in the input parameter range of the GridSearchCV algorithm. In order to find the best combination for the accuracy score, these parameter combinations are used in the implementation of extra gradient boosting classifier model for each of tree cross validation folds of the train set. 2592 fittings are performed in this study and the best parameters are shown in Table 5.

GridSearchCV Output	Parameters After Manual Iterations
'min_child_weight': 1	'min_child_weight': 1
'gamma': 0	'gamma': 0
'subsample': 1.0	'subsample': 1.0
'colsample_bytree': 0.8	'colsample_bytree': 0.8
'max_depth': 10	'max_depth': 25
'eta': 0.3	'eta': 0.3

Table 5: Output of GridSearchCV for extra gradient boosting classifier and parameters obtained after manual iterations.

5.1.2 Results of Finetuned Extra Gradient Boosting Classifier and Random Forest Classifier Models

	Trainset Accuracy	Avg. CV Score of Trainset	Testset Accuracy	Avg. CV Score of Testset	AUC for Testset
RandomForestClassifier	98,29%	87,86%	88,73%	86,77%	95,00%
XGBClassifier	99,53%	87,68%	87,93%	86,05%	95,00%

Table 8: Train and test set accuracies, average values of five different cross validation scores for train and test set, AUC values for test set predictions of fine-tuned extra gradient boosting classifier and random forest classifier models.

Table 8 shows the test and train set accuracy values of the random forest classifier and extra gradient boosting classifier models implementing the parameters obtained by the hyperparameter tuning process. The average accuracy values are obtained from five different cross validation processes separately for the train and test sets. The area under curve (AUC) values of the receiver operator characteristic (ROC) graphs for the test set predictions are also shown in the Table 8. As a result of the hyperparameter tuning study, a slight increase in the test set accuracy values is observed for the both models. On the other hand, the train set accuracy is significantly increased to 99.53% compared to the initial model of extra gradient boosting classifier. In the cross validation process performed on the train set, very close accuracy scores with an average of 87.68% is obtained from the extra gradient boosting classifier model. As a result, there is no overfitting for the extra gradient boosting classifier model.

6. CONCLUSION

The purpose of this study is to determine the reservation details that have an effect on the reservation cancellations and to implement a classification model that can provide the net accommodation amount to be realized within a certain time period by predicting the reservation requests that may be cancelled.

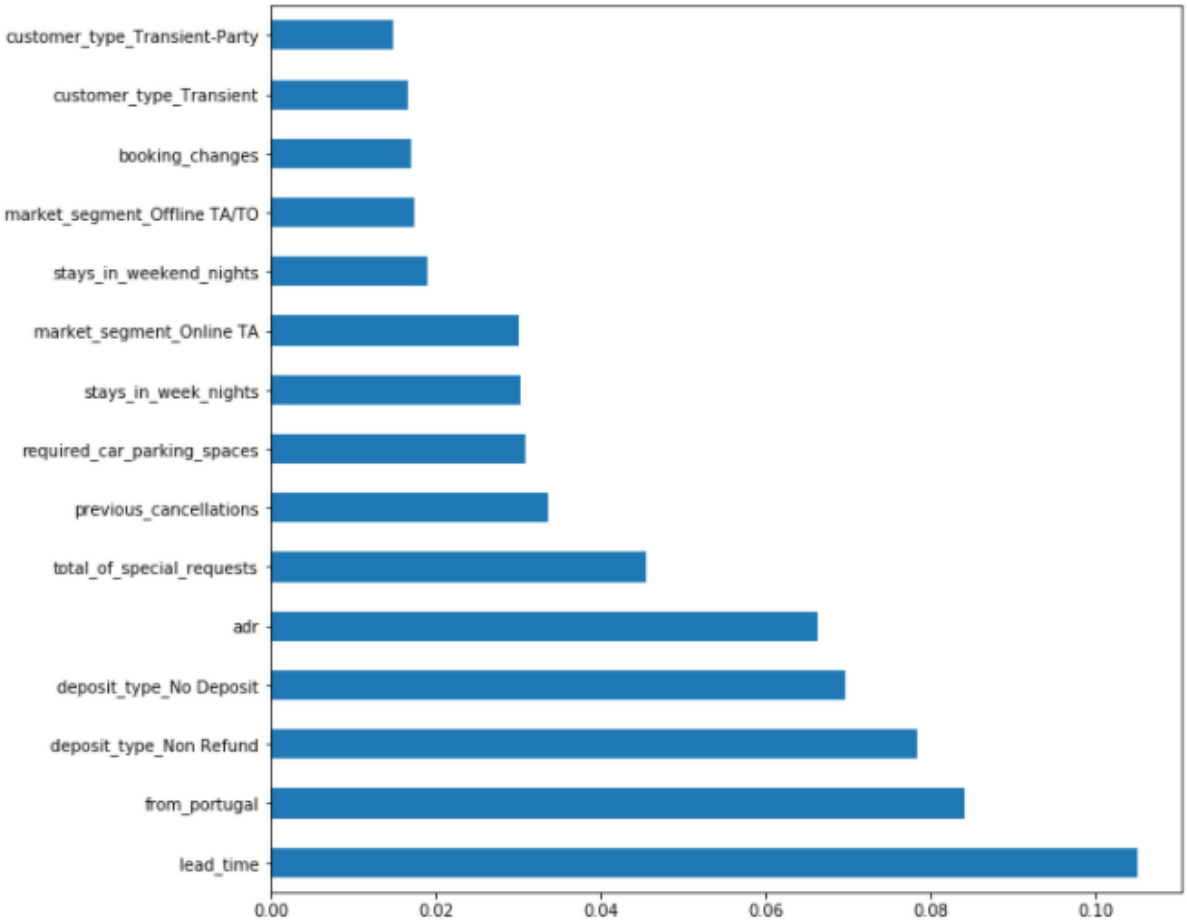

Figure 10: The important features for random forest classifier.

Among the 145 features obtained after implementing preprocessing phase, 15 features that are the most important ones for the random forest classifier model are shown in Figure 10. Accordingly, reservation details such as 'lead_time', 'from_portugal', 'deposit_type', 'adr' and 'total_of_special_requests' have more ability to explain cancellations than other features.

Expanding the dataset for further studies may increase the accuracy of the model created. For example, the ‘is_family’ feature is created additionally to contribute to the model in terms of accuracy and it is shown that the model accuracy is increased. Therefore, obtaining this reservation detail from the customer while creating the requests may increase the accuracy of further models.

Hopefully, this study with other increasing numbers of studies in this subject contributes to improve detection of reservation cancellations so that both hotels and websites where hotel reservations are made can benefit from accurate demand planning, resource allocation and therefore cost savings.

7. REFERENCES

António, N., Almeida, A. D., and Nunes, L. (2019). Predictive models for hotel booking cancellation: A semi-automated analysis of the literature. *Tourism & Management Studies*, 15(1), 7-21.

Antonio, N., Almeida, A. D., and Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, 13(2), 25-39.

Benítez-Aurioles, B. (2018). Why are flexible booking policies priced negatively? *Tourism Management*, 67, 312–325.

Chen, C. (2016). Cancellation policies in the hotel, airline and restaurant industries. *Journal of Revenue and Pricing Management*, 15(3-4), 270-275.

Chen, C., Schwartz, Z., and Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*, 30(1), 129-135.

Chew, E. Y. T., and Jahari, S. A. (2014). Destination image as a mediator between perceived risks and revisit intention: A case of post-disaster Japan. *Tourism Management*, 40(1), 382-393.

Cheyne, J., Downes, M., and Legg, S. (2006) Travel agent vs. internet: what influences travel consumer choices? *Journal of Vacation Marketing*, 12(1), 41-57.

Dolnicar, S., and Laesser, C. (2007). Travel agency marketing strategy: Insights from Switzerland. *Journal of Travel Research*, 46(2), 133-146.

Fesenmaier, D. R., and Jeng, J-M. (2000). Assessing structure in the pleasure trip planning process. *Tourism Analysis*, 5(1), 13-27.

Garrow, L. A., and Koppelman, F. S. (2004). Predicting air travelers' no-show and standby behavior using passenger and directional itinerary information. *Journal of Air Transport Management*, 10(6), 401-411.

Kimes, S. E., and Wirtz, J. (2003). Has Revenue Management become Acceptable? *Journal of Service Research*, 6(2), 125-135.

Law, R., Leung, K., and Wong, J. (2004). The impact of the internet on travel agencies. *International Journal of Contemporary Hospitality Management*, 16 (2), 100-107.

Lemke, C., Riedel, S., and Gabrys, B. (2013). Evolving forecast combination structures for airline revenue management. *Journal of Revenue and Pricing Management*, 12(3), 221-234.

Mehrotra, R., and Ruttley, J. (2006). Revenue management (seconded.). Washington, DC, USA: American Hotel & Lodging Association (AHLA).

Morales, D. R., and Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, 202(2), 554-562.

Mostipak, Jesse. "Hotel Booking Demand." Kaggle, 13 Feb. 2020, www.kaggle.com/jessemostipak/hotel-booking-demand.

Talluri, K. T., and Ryzin, G. V. (2009). The theory and practice of revenue management. New York: Springer.

Tse, S. M. T., and Poon, Y. T. (2017). Modeling no-shows, cancellations, overbooking, and walk-ins in restaurant revenue management. *Journal of Foodservice Business Research*, 20(2), 127- 145.