Predicting Cancellations

Machine Learning in Hotel Data

Name: Chitraang Nayyar
Student No.: 501095666
Supervisor: Dr. Ceni Babaoglu
Date: October 31st, 2022

Ryerson University

Table of Contents

Abstract

Big data is transforming industries everywhere. Hotels and other facilities that are part of the hospitality industry are also looking to embrace big data applications to transform and personalize their customer experience, and to increase their revenues. Hotels leverage big data to customize guest experience, charge optimal prices and develop ways to manage their customers. The biggest challenge for any hotel is to sell the right room to the right customer at the right time and for the right price. Furthermore, given last minute cancellations and "no-shows", there is an additional loss of revenue since the capacity allocation for a hotel is no longer optimal. In my project, I will highlight how machine learning can redefine hospitality industry's revenue management by reducing cancellations.

Hotel booking cancellations have a substantial impact on any hotel's revenue. We will use data science and machine learning to analyze guest data and booking behaviour patterns, and to devise a strategy for hotel revenue management. I will draw upon the dataset used by Nuno Antonio, Ana Almeida, and Luis Nunes' in their co-authored article 'Hotel Booking Demand Datasets' published in Data in Brief (2019) (https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand). Our model aims to understand all the booking related information in order to predict if the guest has a higher chance of cancellation. We are going to deploy the classic classification models such as logistic regression and decision trees using Python to classify a hotel booking's likelihood to be canceled. Techniques will include Data Cleaning, Exploratory Analysis, Descriptive Analysis, Feature Engineering, Modelling to answer the following research questions.

Introduction

Tourism is one of the world's most popular industries and its importance to the global economy is unquestionable. Online travel revenue continues to grow year by year, and in 2019, digital transactions related to online travel reached $755.4 billion. In terms of economic development, the tourism industry generates income to nations through the consumption of goods and services by tourists, taxes, development of enterprises and employment opportunities, among others. However, the environment in which tourism develops its activities is forced to face constant uncertainty. In this sense, the tourism industry is highly sensitive to several external factors that can have a significant impact on revenue, such as political instability, weather, and natural disasters etc. The hotel business operates 24 hours a day, 7 days a week, without exception. People travel around the world for many reasons. Therefore, the customer needs to reserve a room.

Since hotels have a fixed inventory and sell a perishable "product", to make the right room available to the right guest, at the right time, hotels accept bookings in advance. Bookings represent a contract between a customer and the hotel (Talluri & Van Ryzin, 2004). This contract gives the customer the right to use the service in the future at a settled price, usually with an option to cancel the contract prior to the service provision.

Although advanced bookings are considered the leading predictor of a hotel's forecast performance (Smith, Parsa, Bujisic, & van der Rest, 2015), this option to cancel the service puts the risk on the hotel, as the hotel has to guarantee rooms to customers who honor their bookings but, at the same time, has to bear with the cost and the risk for a customer canceling a booking or a no-show (Talluri & Van Ryzin, 2004). To mitigate the difficulties caused by cancellations, hotels implement rigid cancellation policies and overbooking strategies (Mehrotra & Ruttley, 2006). On the other hand, strict cancellation policies, especially non-refundable policies, not only reduce the number of bookings, but can also reduce revenue by applying steep discounts (Smith & Bujisic, 2015).The methods, (Kumari & Srivastava, 2017) completed a taxonomy for binary classification. Therefore, in the hospitality industry booking cancellations play a vital role as management decisions are clearly influenced by customer demand, as both bookings and cancellations are major components of revenue management. One of the most important tools in Revenue Management System is its predictive power. Without accurate forecasting capabilities, a revenue management system's rate and availability recommendations can be highly unreliable. In fact, an accurate booking cancellation forecast is paramount to determining net demand. Hotels try to implement overbooking strategies to avoid losing revenue, which is understandable because of the increase in cancellations.

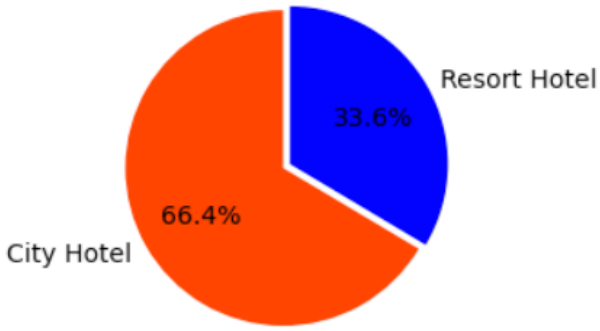
This work uses advanced data science techniques to summarize current research on predicting/predicting model development of booking cancellations in the hotel industry and identifies the main themes of booking cancellation research. Using this data set, we can do the following analyzes:

- Where do the guests come from?
- When is the best time of year to book a hotel room?
- How much do guests pay for a room per night?
- What's the earned revenue and the lost revenue?
- How long do people stay at the hotels?
- What is the optimal length of stay to achieve the best daily price?
- Which month has the highest number of cancellations?

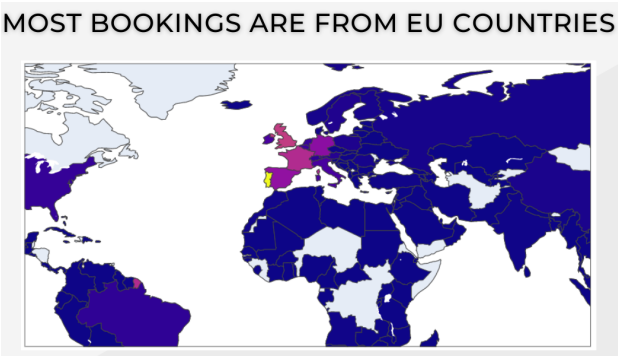
Data Description

I will draw upon the dataset used by **Nuno Antonio, Ana Almeida, and Luis Nunes' in their co-authored article 'Hotel Booking Demand Datasets' published in Data in Brief (2019) (https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand).** The bookings are made from July 1st 2015 to August 31st 2017.The dataset includes 32 features/variables before cleaning. The features are a mixture of 15 categorical variables and 17 numerical variables.

Integer	*LeadTime, ArrivalDateYear, ArrivalDateWeekNumber, ArrivalDateDayOfMonth, StaysInWeekendNights, StaysInWeekNights, Adults, Babies, IsRepeatedGuest, PreviousCancellations, PreviousBookingsNotCanceled, BookingChanges, DaysInWaitingList, RequiredCarParkingSpaces, TotalOfSpecialRequests*
Float	*Children, Agent, Company, ADR*
Object	*ArrivalDateMonth, Meal, Country, MarketSegment, DistributionChannel, ReservedRoomType, AssignedRoomType, DepositType, CustomerType, ReservationStatusDate*



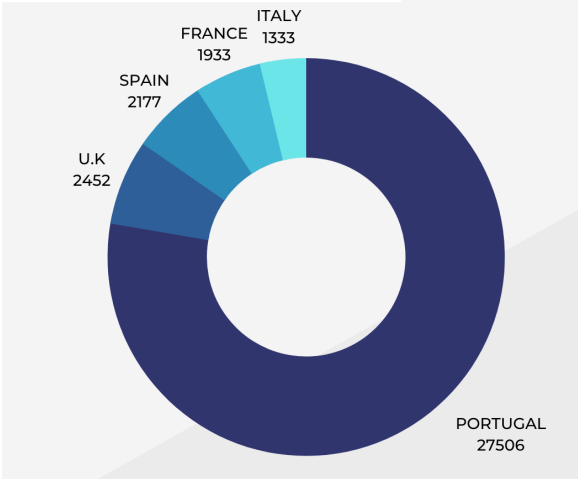
The dataset includes hotel booking information of two hotels: a city hotel and resort hotel in Portugal with 119,390 observations. City Hotel bookings have shown higher bookings (79,163) than Resort Hotel is (40,047). At this stage, data visualization is done to understand the data and clean the data from the lost data or delete the problematic features to produce a better and more general machine learning model.



Our research showed 48,590 bookings were made from the home country, Portugal from July 1st 2015 to August 31st 2017. Followed by United Kingdom and France at 12,129 and 10,415. Similarly, most cancellations recorded from within Portugal 27,506, followed by United Kingdom 2,452 and Spain 2,177.

Data preprocessing is used to clean and manipulate dataset for further analysis. In addition, different cases such as detecting outliers, missing values and purifying data were scrutinized in order to build well generalized model.

The null values in each feature stand for similar meaning, we just replace the nulls with particular integer or text instead. As a result, we removed the company column that leads to 94% missing values.



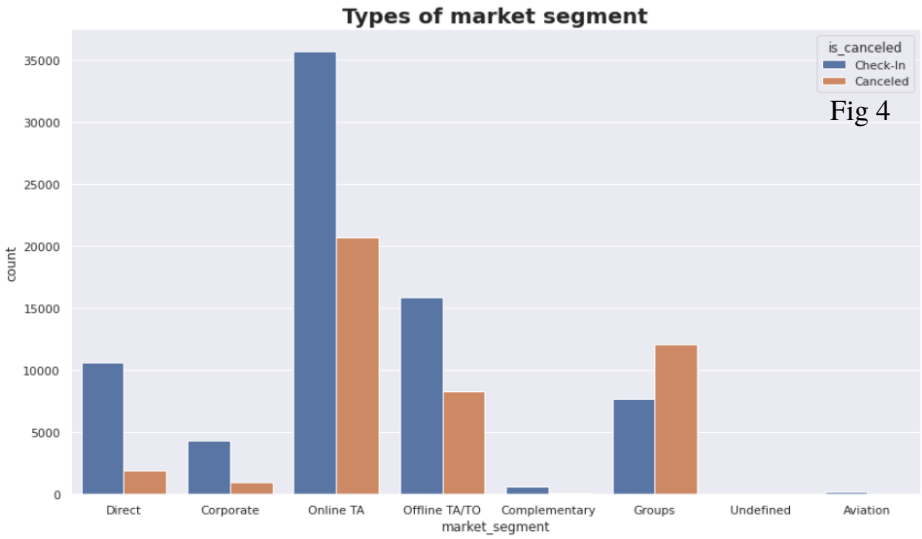


Fig 4

Visualization of Guests and the Market Segments

Fig 4 shows most bookings are through Travel Agents/Tour Operators and similarly the higher cancellations are reported through Travel Agents/Tour Operators.

The average cancellation rate is around 40% every quarter and there is no seasonality in the ratio of the cancellation rate, while the resort hotels there is seasonality , where the cancellation rate occurs at its peak in the third quarter (July until September) and also low in the first and fourth quarter. In the winter months, a smaller number of customers come, so when we look at the cancellation rates it is quite normal for a smaller number to appear in the winter months. We can observe that these months, city hotel cancellation rates are almost the same as other months even in winter. Because overall cancellation rates in the winter months are low, then the cancellation rates of the resort hotels are low in these months. We can conclude that the possibility of canceling summer resorts in winter is very small.

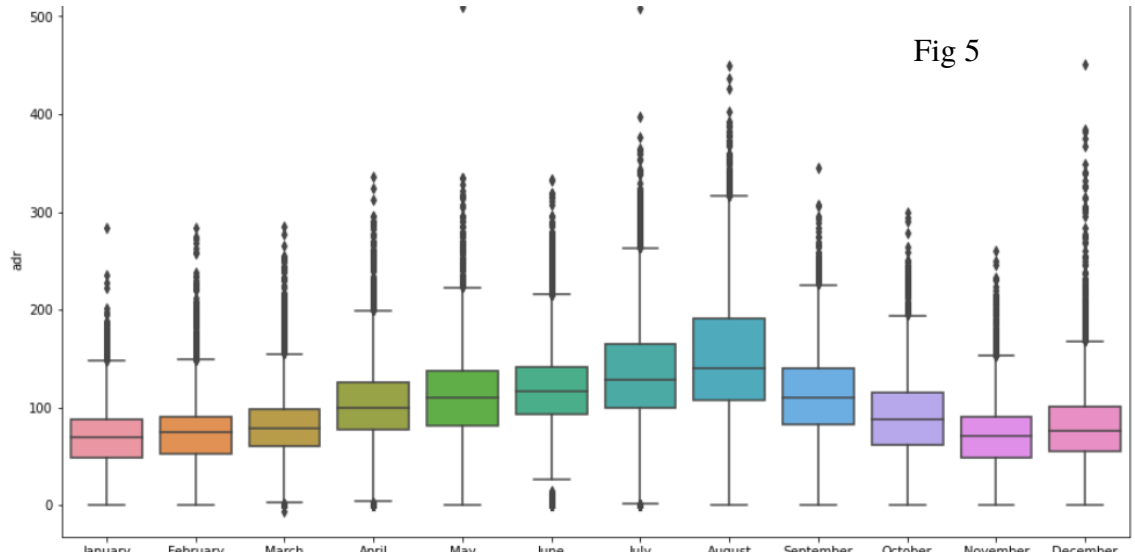


Fig 5

Visualization of ADR by the month

As shown in Fig 6 and Fig 7, most cancellations for both City and Resort Hotel are received on the day of check-in. Considering this cancellation forecast, this trend puts the risk on the hotel, leading to a loss of revenue as that room could have been allotted to a new guest.

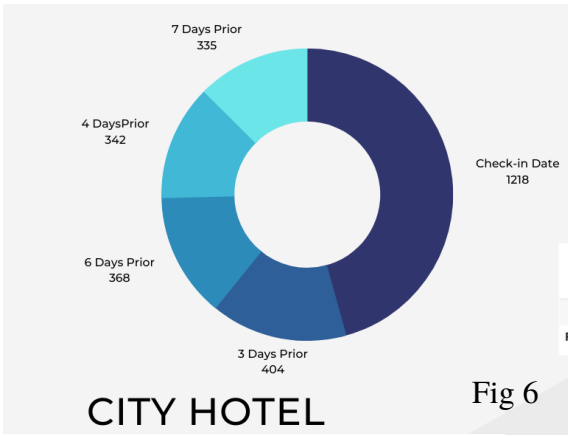


Fig 6

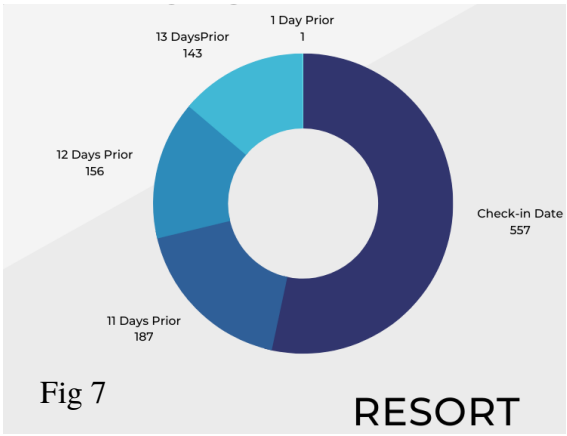
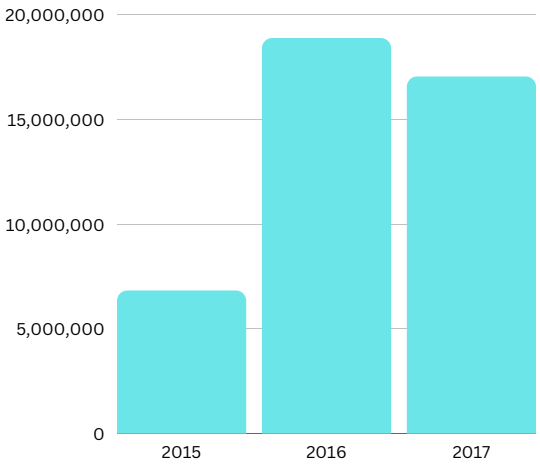
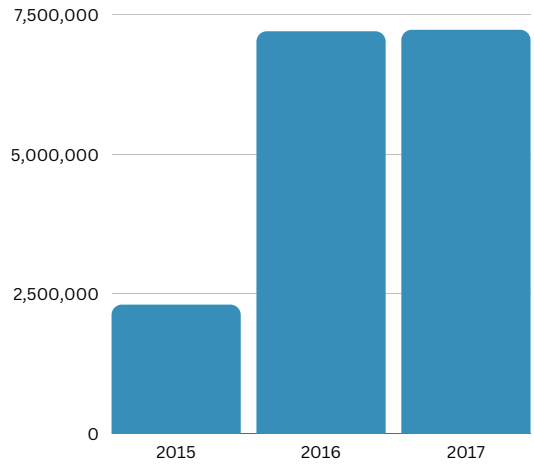


Fig 7



Revenue Earned Each Year (EURO)



Revenue Lost Each Year (EURO)

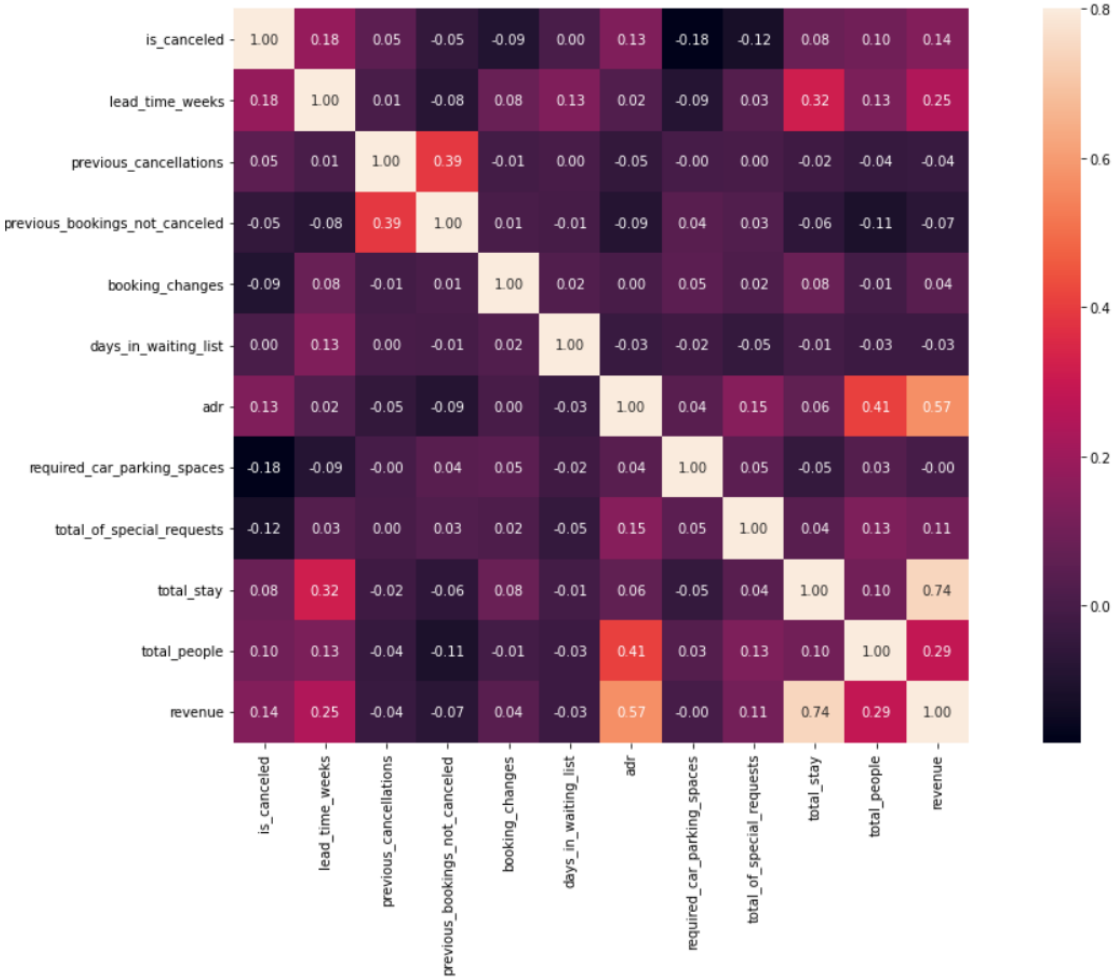


Fig 8 Correlation Matrix

The correlation matrix is a table showing the correlation coefficients between the variables. In the table each cell shows the relationship between the two variables. Correlation matrix is used to summarize data, as an input into a more advanced analysis, or as a diagnostic for advanced analyses. We presented the correlation matrix for hotel demand data set in Fig 8. We observe the attribute “is_canceled”, is strongly correlated with the status of the revenue, the total stay, Average Daily Rate(ADR) and lead time.

Modelling

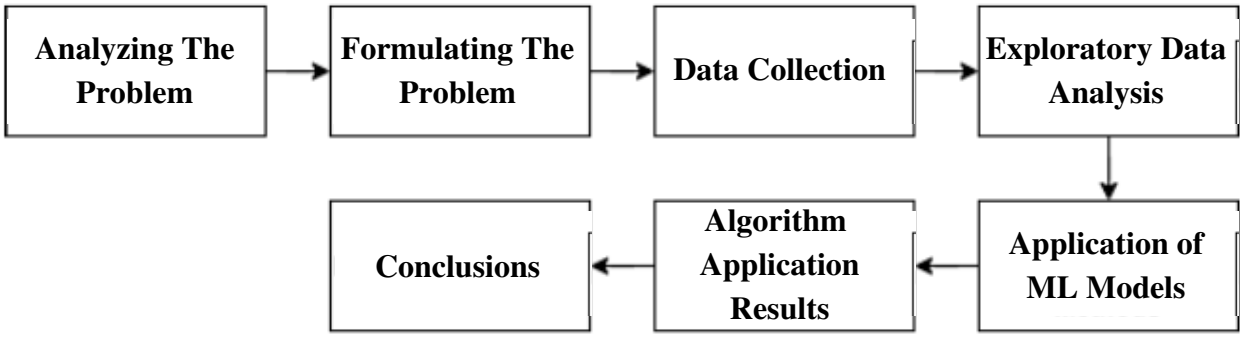


Fig. Visualization of Research Stages

In this phase, the selection and application of variations in machine learning models are carried out, so that the best model is obtained. We used these popular machine learning algorithms: Logistic Regression, K- Nearest Neighbour, Decision Tree Classifier and Extra Tree Classifier .

- **Logistic Regression :** The baseline model is set as the benchmark for the performance comparison. Since this is binary classification problem and logistic regression can map all data points into a value between 0 and 1 by the sigmoid function, then it's used as the baseline model.

- **Decision Tree :** The decision tree can break down the problem like binary classification into a bunch of subsets with homogenous values. The splitting procedure is helpful to show the importance of each feature and thus provide us some insights. After all, a big advantage of decision tree is its interpretability which is close to human-being's decision making process.

- **K-Nearest Neighbours :** The K-Nearest Neighbours is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. This algorithm assumes that similar things exist in close proximity, calculating the distance between points on a graph. The algorithm is versatile. It can be used for classification or regression.

- **Extra Trees :** Extra trees is also ensembled by decision trees. However different from optimal splitting points, extra trees uses the whole original sample and split nodes by random. In other words, extra trees has a higher degree of randomness but also keeps optimization. Therefore, extra trees may execute faster.

Model	Accuracy Score
Logistic Regression	**0.813494**
K - Nearest Neighbours	**0.892794**
Decision Tree	**0.949445**
Extra Tree	**0.952633**

Fig 9

For hotel reservation data, during the training phase logistic regression showed the weakest performance of the classification (Fig 9). K- Nearest Neighbours classifier had better classification accuracy. Decision tree and Extra Tree classifiers showed the best accuracy results. For the mentioned classifiers in training phase, we used grid search during cross validation to find the classifiers with the best parameters.

Accuracy can be represented as a proportion of correctly classified cases:

$$\textbf{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

We have shown in Fig 10 the results obtained during the training and test phases using cross validation for all classifiers. In this table:

- **True Positive (TP) means cancelled transaction correctly diagnosed as cancelled transactions,**
- **False positive (FP) means not cancelled transactions incorrectly identified as cancelled transactions,**
- **True negative (TN) means not cancelled transactions correctly identified as not cancelled transactions,**
- **False negative (FN) means cancelled transactions incorrectly identified as not cancelled transactions.**

Detailed results in Fig 10 allows predictions of booking cancellations. In this way we can see for which class our classifier works better.

Classifiers	TN	FP	FN	TP
Logistic Regression	21179	1263	5502	7819
K-Nearest Neighbours	21606	836	3113	10208
Decision Tree	21587	855	892	12429
Extra Tree	22220	222	1595	11726

Fig 10

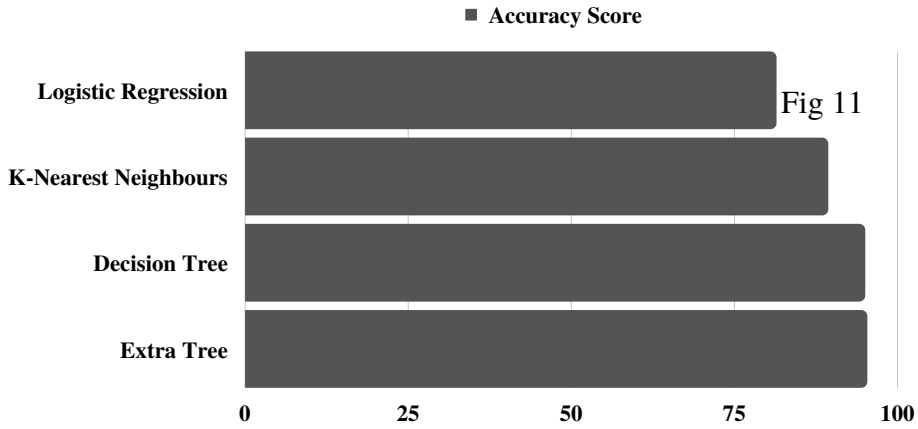


Fig 11

We showed in Fig 11, the training accuracy of the classification and for the classifiers with the best parameters.

Conclusion

Predicting cancellations of hotel bookings is important in order to minimize the loss of revenue to the hotel. The proposed methods, Decision Tree and Extra Tree classifiers allow hotel managers to take actions on bookings identified as “potentially going to be canceled”, but also to produce more precise demand forecasts. A noteworthy suggestion is for the hotel to send out promotional emails to clients before the arrival date. This has the potential to increase revenues as most cancellations are received on the day of check-in. Booking cancellations models also allow hotel managers to implement less rigid cancellation policies, without increasing uncertainty. This has the potential to translate into more sales, since less rigid cancellation policies generate more bookings. By using the stages in Decision Tree and Extra Tree, the most appropriate features for the machine learning model are built, the best machine learning model is Extra Tree, with an accuracy value of 0.9526 and the time difference between booking date and arrival date is the most influential feature to predict the cancellation rate of a hotel booking. as Talluri & Van Ryzin (2004) said “this combination of science and technology applied to age-old demand management is the hallmark of modern revenue management” .

References

- Antonio N., Almeida A. and Nunes L. 2019. "Hotel booking demand datasets". *Data in Brief*, Elsevier, Volume 22, pp. 41-49, February 2019

- Antonio N., de Almeida A., and Nunes L. 2017. "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model". *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, 2017, pp. 1049-1054. doi: 10.1109/ICMLA.2017.00-11.

- Antonio N., de Almeida A., and Nunes L. 2017. "Predicting hotel booking cancellations to decrease uncertainty and increase revenue", *Tourism & Management Studies*, 13(2), 2017, 25-39 DOI: 10.18089/tms.2017.13203.

- K. T. Talluri and G. Van Ryzin, The theory and practice of revenue management, New York, NY:Springer, 2004.

- S. J. Smith, H. G. Parsa, M. Bujisic and J.-P. van der Rest, "Hotel cancelation policies distributive and procedural fairness and consumer patronage: A study of the lodging industry", J. Travel Tour. Mark, vol. 32, no. 7, pp. 886-906, Oct. 2015.

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

- Roshan Kumari and Saurabh Kr Srivastava. Machine learning: A review on binary classification. International Journal of Computer Applications, 160(7), 2017.

- Yueqian Zhang. Forecasting Hotel Demand Using Machine Learning Approaches. PhD thesis, 08 2019.