**Project Title**

**ANALYSIS AND PREDICTION OF HOTEL BOOKING USING MACHINE LEARNING**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfilment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**SONU M**

**PREETHY A**

**ANJALI GJ**

**RASEENA S**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Nov 2022**

**List of Figures**

| **Fig No** | **Figure** | **Page Number** |
| --- | --- | --- |
| 4.1 | Flow Diagram of Procedure | 15 |
| 4.2 | Pie Chart of Hotel types and Target column | 20 |
| 4.3 | Room Type Distributions | 21 |
| 4.4 | Percentage Cancellations Hotel Wise | 22 |
| 4.5 | Density Plot of Waiting list days and lead time | 23 |
| 4.6 | Density Plot of ADR | 23 |
| 5.1 | Heat Map of target column vs features | 29 |
| 8.1 | Fs scores of all features in descending order | 47 |

**List of Tables**

| **Table No** | **Table** | **Page Number** |
| --- | --- | --- |
| 4.1 | Dataset Column Information | 16 |
| 6.1 | Accuracy scores of Models | 37 |
| 6.2 | Feature Selection Parameters | 39 |
| 6.3 | Accuracy scores after feature selection | 40 |
| 8.1 | Accuracy scores of Models | 45 |
| 8.2 | K fold Cv scores of Models | 46 |
| 8.3 | .Stratified K fold Cv scores of Models | 46 |
| 8.4 | Accuracy scores after feature selection | 48 |
| 8.5 | K fold CV scores of Models-10 Features | 48 |
| 8.6 | Stratified K fold CV scores of Models-10 Features | 49 |

**List of Abbreviations**

| ADABoost | Adaptive Boosting |
| --- | --- |
| ADR | Average Daily Rate |
| ANOVA | Analysis of variance |
| API | Application Programming Interface |
| BB | BB – Bed & Breakfast; |
| CV | Cross Validation |
| FB | Full board (breakfast, lunch and dinner) country |
| FS | Feature Selection |
| H1 | Resort hotel |
| H2 | City hotel |
| HB | Half board (breakfast and one other meal – usually dinner) |
| KNN | K-nearest neighbors |
| LGBM | Light Gradient Boosting Machine |
| PCA | Principal component analysis |
| TA | Travel agents |
| TO | Tour operators |
| XGB | Extreme Gradient Boosting Algorithm |

**Table of Contents**

| **Sl no** | **Content** | **Page Number** |
| --- | --- | --- |
|  | **Abstract** | 7 |
| 1  1.1  1.2 | **Problem Definition**  **Overview**  **Problem Statement** | 8 |
| 2 | **Introduction** | 9 |
| 3 | **Literature Survey** | 10 |
| 4  4.1  4.1.1  4.2  4.2.1  4.2.2 | **Dataset and Exploratory Data Analysis**  Dataset  Column information  Exploratory Data Analysis  Univariate Analysis  Bivariate Analysis | 15 |
| 5  5.1  5.2  5.3  5.4  5.5  5.6  5.7  5.8  5.9 | **Pre-Processing**  Duplicate Value Handling  Missing Value Handling  Converting Data Types  Feature Reduction  Feature Extraction  Encoding  Outlier handling  Scaling on the Dataset  Heat Map Analysis | 25 |
| 6  6.1  6.2  6.3  6.4  6.5  6.6  6.7  6.8  6.9  6.10  6.11  6.11.1  6.12 | **Model Building**  Model 1-Decision Tree Algorithm  Model 2-Random Forest algorithm  Model 3.-Logistic Regression  Model 4.-KNN Algorithm  Model 5.-Gradient Boosting Algorithm  Model 6-XGB Classifier  Model 7-LGBM Classifier  Model 8-Extra Tree Classifier  Model 9-AdaBoost Classifier  Cross Validation  Feature Selection  Select-KBest Feature Selection Library  Hyper parametric Tuning of the models | 30 |
| 7 | **Web Deployment** | 43 |
| 8 | **Result** | 45 |
| 9 | **Conclusion** | 50 |
|  | **References** | 51 |

**Abstract**

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, number of available parking spaces, average daily rate of transactions by a customer etc. The dataset comprises 36 variables and 119390 observations. The dataset comprehends bookings made between the year 2015 and 2017 in Portugal, including bookings that effectively arrived and bookings that were cancelled. This project aims to demonstrate how the predictive analysis of the model can contribute to booking cancellation forecasting, by implementing various Algorithms like Logistic, SVM models, KNN, Random Forest, Decision Tree, etc. It prevents the hotel as well as Tourists from poor dealing of rooms. User/Customer has to enter certain fields by which this model detects his prediction about the cancellation. Furthermore we will be analysing some key metrics for hotel bookings and also we will be looking through the data to analyse patterns associated with each segment. Using the results from the above analysis, businesses can make key decisions regarding the customer experience they desire to deliver

**1. Problem Definition**

**1.1 Overview**

Hotels are definitely one of the fastest-growing sectors in the tourism sector and it is truly justified as accommodation is the key part in the development of any country or region's tourism. Tourism and Hotel Industry always go hand in hand and the presence of enough hotels adds value to a region's economy.This project aims to create meaningful estimators from the data set we have and to perform Exploratory Data Analysis so that if anyone who wishes to proceed with the Machine Learning Model, can do so.When running a successful and demanding hospitality business, most hotel owners like a hotel that is running at full capacity and bringing in sizeable revenue. Most of the time hotel booking cancellations can be hurtful to business owners; although sometimes there are genuine reasons for guests to do so. These last-minute cancellations can result in lost revenue unless some measures are undertaken to mitigate the loss. The purpose of this project is to analyse Hotel Bookings data, investigate cancellations, and their underlying patterns; and suggest measures that can be implemented to reduce cancellations and secure revenue.This Hotel Booking cancellation model can be useful not only for the hotels owners but for the hotel vacationers.

**1.2 Problem Statement**

Hotel industry is a very volatile industry and the bookings depend on a variety of factors such as type of hotels, seasonality, days of week and many more. This makes analysing the patterns available in the past data more important, to help the hotels plan well. Booking cancellation has a significant effect on revenue. By combining data science tools and capabilities with human judgement and interpretation, this project aims to demonstrate how the predictive analysis of the model can contribute to synthesising and predicting booking cancellation forecasting.

Furthermore we will be analysing some key metrics for hotel bookings like:Where do the guests come from, How much do guests pay for a room per night,How does the price per night vary over the year, Which are the busiest month,How long do people stay at the hotels,Bookings by market segment, How many bookings were cancelled, Which month has the highest number of cancellations, Repeated guest effect on cancellations,The number of nights spent at hotels, Hotel type with more time spent,Effects of deposit on cancellations by segments, Relationship of lead time with cancellation,Monthly customers and cancellations.Finally, we will also try to predict whether guest cancel reservation and use some explanation methods to analyse the reasons of customer behaviour.

**2. Introduction**

Using the historical data, hotels can perform various campaigns to boost the business. We can use the patterns to predict the future bookings using time series or decision trees.This dataset contains information on records for client stays at hotels. More specifically, it contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.Hotels plays an important role for any person or traveller who are travelling from one destination to another. It plays an important role for tourists whether the tourist is local or international. Hotel provides many best services to the customer such as parking area, food, room service and also it provides services that are offered by the customer. By providing these services Hotels take valuable feedback from the customers. By these feedbacks Hotels maintain their reputation in the city/area. If the services are poor, the bookings of that hotel are low and if the services are awesome then bookings are high.

This dataset contains booking information for a city hotel and a resort hotel in Portugal, and includes information such as when the booking was made, length of stay, the number of adults, children, babies, the number of available parking spaces, chosen meals, price etc. There are 119390 observations and 36 features.This data set does not contain any duplicate values.The columns name,age,email id,phone number and company mostly may not have an impact on analysis and it contains a high number of null values so they can be dropped .There are null values in agent,company and country columns. They are to be filled as part of preprocessing. The dataset also contains outliers in continuous variables columns so they are to be handled. Only after these steps the model creation can be done in the dataset.

In this project the dataset which we are using contains both International and Local Hotel data. Here we use many popular Machine Learning Algorithms like Decision Tree, Random Forest, KNN, Logistic Regression, etc. to predict the cancellation chances. After that web hosting will be done to predict hotel booking cancellation status.

**3. Literature Survey**

A detailed literature study has been conducted for the project work. These literature reviews provide an overview of past and current research relating to similar work.

In the paper ‘Application of Machine Learning in the Hotel Industry: A Critical Review’ by **Dr.Eid Alotaibi. D [1]** gives insights on the role of ML and its integrated technologies in the hotel industry.The study found that machine learning is helpful in demand forecasting, price forecasting, booking cancellation prediction, financial efficiency, and work efficiency. The machine learning algorithms outperform in the forecast accuracy against the statistical models.In the research paper he conducts exploratory analysis to identify the extent of scientific community knowledge and awareness on machine learning in the hotel industry. In his paper he suggested that the researchers are required to investigate numerous areas that have been missed out of machine learning. The three most important areas entitled to focus in future research for machine learning design are the services quality, revenue management, and management strategy. All three are interlinked and management strategies control the other two. The hotel industry will improve service quality and generate more revenue by strategically taking automation decisions.

In the research paper ‘Hotel Booking Prediction using Machine Learning’ by **Pranav Kumar and Sarthak Sharma [2]** aims to demonstrate how the predictive analysis of the model can contribute to synthesising and predicting booking cancellation forecasting. In the research work, by implementing Various Algorithms like Logistic, KNN, Random Forest, Decision Tree, etc. to classify the data and also use Evaluation Matrix to separate categorical data in particular type, they made prediction up to the desired level which prevented the hotel as well as Tourists from poor dealing of room.Tuned Random Forest Has the Best Accuracy Among All Algorithm That they tried from all the evaluation matrix. This model enables hotel managers to mitigate revenue loss derived from booking cancellations and to mitigate the risks associated with overbooking (relocation costs, cash or service compensations, and, particularly important today, social reputation costs). Their work also allows hotel managers to implement less rigid cancellation policies, without increasing uncertainty. This research paper has the potential to reduce cancellation policies which generate more bookings.

To reduce the cancellation effect a machine learning based system prototype was developed by **Nuno Antonio, Ann de Almeida, Luis Nunes [3]** in their research work ‘Predicting Hotel Bookings Cancellation with machine learning classification model’. It makes use of the hotel’s Property Management Systems data and trains a classification model every day to predict which bookings are “likely to cancel” and with that calculate net demand. The prototype was deployed in a production environment in two hotels, by enforcing A/B testing, which also enables the measurement of the impact of actions taken to act upon bookings predicted as “likely to cancel”. Results indicate good prototype performance and provide important indications for research progress whilst evidencing that bookings contacted by hotels cancel less than bookings not contacted. The Work presented good results for both hotels, with an Accuracy above 0.84. This shows that there is some space for improvement. In terms of future research, these models could benefit from the introduction of features from other data sources related to factors that affect customers booking/cancellation decisions, like competitors’ prices, competitors’ social reputation, weather, among others. Another feature that has the potential to improve model accuracy, taking in consideration the impact actions can have on customers’ decision not to cancel, is a feature that identifies if an action to avoid cancellation was already taken on the booking. Given sufficient time, the system can have the potential to generate a labelled database with the actions made in each booking to avoid cancellation. The database they worked on could also be used to develop another model which, in combination with this model, suggests which actions should be appropriate to take in each identified booking.

In the research paper ‘An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellation Authors’ by **Nuno Antonio , Ana de Almeida, Luis Nunes [4]** Research has shown that with today’s computational power and advanced machine learning algorithms it is possible to build models to predict bookings cancellation likelihood. However they suggested that effectiveness of these models has never been evaluated in a real environment..A prototype was built and deployed in two hotels. The prototype, based on an automated machine learning system designed to learn continuously, led to two important research contributions. First, the development of a training method and weighting mechanism designed to capture changes in cancellations patterns over time and learn from previous days’ predictions hits and errors. Second, the creation of a new measure – Minimum Frequency – to measure the precision of predictions over time. From a business standpoint, the prototype demonstrated its effectiveness, with results exceeding 84% in accuracy, 82% in precision, and 88% in Area Under the Curve (AUC). The system allowed hotels to predict their net demand and thus make better decisions about which bookings to accept and reject, what prices to make, and how many rooms to oversell. The systematic prediction of bookings with high probability of being cancelled allowed hotels to reduce cancellations by 37 percentage points by acting to avoid their cancellation.This study highlights how a service-oriented decision support system, based on an automated machine learning model, designed in accordance to DSR to address an unsolved problem in a unique and innovative manner, can be constructed and implemented.Future research can explore the predictive power of features not only to better understand cancellation drivers but also to use this knowledge to improve cancellation policies.

In the research paper ‘Predicting hotel booking cancellations to decrease uncertainty and increase revenue’ by **Antonio, N., de Almeida, A. and Nunes, L [5],.**Using data sets from four resort hotels and addressing booking cancellation prediction as a classification problem in the scope of data science, authors demonstrate that it is possible to build models for predicting booking cancellations with accuracy results in excess of 90%. Results allow hotel managers to accurately predict net demand and build better forecasts, improve cancellation policies, define better overbooking tactics and thus use more assertive pricing and inventory allocation strategies.They were able to Identify which features contributed to predict a booking cancellation probability. Application of data visualisation and data analytics techniques, together with the application of the mutual information filter, allowed the understanding of a feature's predictive relevance. It was found that different features differ in importance depending on the hotel, and some features are not required for some of the hotels. It was established that, depending on the hotel, around 30 features are enough to build a good prediction model. They suggested that deployment of these predictive models in a production environment, in hotels, with the purpose of executing A/B testing, could contribute to measuring the effect of having previous knowledge of which bookings have a high cancellation probability.

In the research paper ‘Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis’ by **Soo Y. Kim [6] ,**The objective of the paper is to provide an optimal hotel bankruptcy prediction approach to minimise the empirical risk of misclassification and secondly, to investigate the functional characteristics of multivariate discriminant analysis, logistic, artificial neural networks (ANNs), and support vector machine (SVM) models in hotel bankruptcy prediction. The performances were evaluated not only in terms of overall classification and prediction accuracy but also in terms of relative error cost ratios. The results showed that ANN and SVM were very applicable models in bankruptcy prediction with data from Korean hotels. When jointly measuring both type I and type II errors, especially allowing for the greater costs associated with type I errors, however, ANN was more accurate with smaller estimated relative error costs than SVM. Thus, if the objective is to find the best early warning technique that performs accurately with small relative error costs, then, it will be worth considering ANN method for hotel bankruptcy prediction.

In the research paper ‘Using data science to predict hotel booking cancellations’ by A**ntonio, N, de Almeida, A. and Nunes, L [7],** Using Booking cancellations in the hospitality industry not only generate revenue loss and affect pricing and inventory allocation decisions, but they also, in overbooking situations, have the potential to affect the hotel’s online social reputation. By employing data sets from four resort hotels and addressing this issue as a classification problem in the scope of data science, the authors demonstrate that it is possible to build models for predicting booking cancellations with accuracy results in excess of 90%. This research also demonstrates that despite what was alleged by Morales and Wang (2010), it is possible to predict with high accuracy whether a booking will be cancelled. Results allow hotel managers to act on bookings with high cancellation probability and contain the associated revenue losses, produce better net demand forecasts, improve overbooking/cancellation policies, and have more assertive pricing and inventory allocation strategies

In the research paper ‘Prediction of Hotel Booking Cancellation using CRISP-DM’ By **Zharfan Akbar Andriawan *et al* [8],**Online travel sales continue to increase every year. Recorded in 2019, digital transactions related to online travel reached USD 755.4 billion. One of the supports of the travel business is the tourism and hospitality industry. The online reservation system is one of the most attractive solutions in the hospitality industry. Cancellation of hotel bookings or reservations through the online system is currently one of the problems in the hotel management system. When the reservation has been cancelled, the hotel will be harmed. Therefore, predicting whether a booking will be cancelled or not using the help of data science is needed so that the hotel can minimise lost profits. Therefore, by using datasets related to hotel booking requests, a predictive analysis using the CRISP-DM framework is conducted. By first performing some data preparation processes, this study uses a tree-based algorithm to perform the prediction. The experiment produced that Random Forest model has the best value with an accuracy value of 0.8725 and it is found that the time difference between booking is made and arrival time is the most influential feature in predicting the level of cancellation.

In the research paper ‘Prediction of hotel booking cancellation using deep neural network and logistic regression algorithm’ By **Nugroho Adi Putro1, Rendi Septian Widiastuti, Mawadatul Maulidah, Hilman Ferdinandus Pardede [9]** , Booking cancellation is a key aspect of hotel revenue management as it affects the room reservation system. Booking cancellation has a significant effect on revenue which essentially affects request board choices in the inn business. To reduce the cancellation effect, the hotel applies the cancellation model as the key to addressing this problem with the machine learning-based system developed. This study, using data collected from the Kaggle website with the name hotel-booking demand dataset. The research objective was to see the performance of the deep neural network method which has two classification classes, namely cancel and not. Then optimised with optimizers and learning rate. And to see which attribute has the most role in determining the level of accuracy using the Logistic Regression algorithm. The results obtained are the Encoder-Decoder Layer by adamax optimizer which is higher than that of the Decoder Encoder by adadelta optimizer. After adding the learning rate, the adamax accuracy for the encoders and encoders decreased for a learning rate of 0.001. The results of the top three ranks of each neural network after adding the learning rate show that the smaller the learning rate, the higher the accuracy, but we don't know what the optimal value for the learning rate is. By using the Logistic Regression algorithm by eliminating several attributes, the most influential level of accuracy is the state attribute and total\_of\_special\_requests, where accuracy increases when the state attribute is removed because there are 177 variations in these attributes.

In the research paper ‘Predictive models for hotel booking cancellation: a semi-automated analysis of the literature’,**Nuno António,Ana de Almeida, Luis Nunes [10]** In reservation-based industries, an accurate booking cancellation forecast is of foremost importance to estimate demand. By combining data science tools and capabilities with human judgement and interpretation, this paper aims to demonstrate how the semiautomatic analysis of the literature can contribute to synthesising research findings and identify research topics about booking cancellation forecasting. Furthermore, this work aims, by detailing the full experimental procedure of the analysis, to encourage other authors to conduct automated literature analysis as a means to understand current research in their working fields. The data used was obtained through a keyword search in Scopus and Web of Science databases. The methodology presented not only diminishes human bias, but also enhances the fact that data visualisation and text mining techniques facilitate abstraction, expedite analysis, and contribute to the improvement of reviews. Results show that despite the importance of bookings’ cancellation forecast in terms of understanding net demand, improving cancellation, and overbooking policies, further research on the subject is still needed.

**4. Dataset and Exploratory Data Analysis**

**4.1 Dataset**

The Dataset which is used in this project is taken from Kaggle.com. The Data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. This Dataset contains the details of Hotels situated in Local and International places. Most of the Hotel Data is from Portugal. Also, the data contains only two Hotels Name i.e. Resort and City Hotel assuming that the City Hotel means the Hotel Data for the City and Resort Hotel means the Hotels which are Hotel cum Resort. The Following Procedures have to be done on the dataset for analysis and prediction

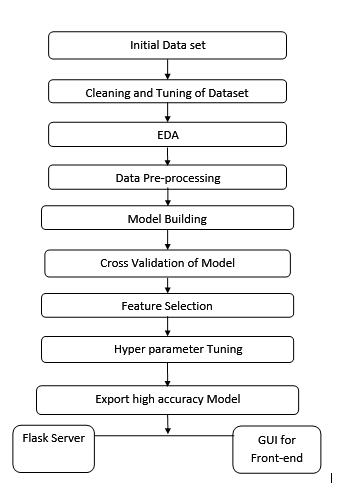


Fig.4.1.Flow Diagram of Procedure

This Dataset contains four types of Data points, these are mentioned below.

1) Market Segments: Offline or Online sources from where the booking is generated and related variables.

2) Hotel & Revenue: City or Resort Hotel, ADR (Average Daily Rate), etc.

3) Customer Related: Variables describing the type of customer based on stay.

4) Cancellation History: Has there been cancellation earlier? etc.

**4.1.1.Column information**

There are 36 columns in the dataset. They are

| **Sl. no** | **Column name** | **Description** |
| --- | --- | --- |
| 1 | Hotel | H1: Resort hotel  H2: City hotel |
| 2 | is\_canceled | 1: Cancelled  0: Not cancelled |
| 3 | lead\_time | No of days that elapsed between entering date of booking into property management system and arrival date |
| 4 | arrival\_date\_year | Year of arrival date (2015-2017) |
| 5 | arrival\_date\_month | Month of arrival date (Jan - Dec) |
| 6 | arrival\_date\_week\_number | Week number of year for arrival date (1-53) |
| 7 | arrival\_date\_day\_of\_month | Day of arrival date |
| 8 | stays\_in\_weekend\_nights | No of weekend nights (Sat/Sun) the guest stayed or booked to stay at the hotel |
| 9 | stays\_in\_week\_nights | No of week nights (Mon - Fri) the guest stayed or booked to stay at the hotel |
| 10 | Adults | Number of Adults |
| 11 | Children | Number of Children |
| 12 | Babies | Number of babies |
| 13 | Meal | Type of meal booked.  Undefined/SC – no meal package;  BB – Bed & Breakfast;  HB – Half board (breakfast and one other meal – usually dinner);  FB – Full board (breakfast, lunch and dinner) country |
| 14 | Country | Country of the customer |
| 15 | market\_segment | (a group of people who share one or more common characteristics, lumped together for marketing purposes)  TA: Travel agents  TO: Tour operators |
| 16 | distribution\_channel | (A distribution channel is a chain of businesses or intermediaries through which a good or service passes until it reaches the final buyer or the end consumer)  TA: Travel agents  TO: Tour operators |
| 17 | is\_repeated\_guest | (value indicating if the booking name was from repeated guest)  1: Yes  0: No |
| 18 | previous\_cancellations | Number of previous bookings that were cancelled by the customer prior to the current booking |
| 19 | previous\_bookings\_not\_canceled | Number of previous bookings not cancelled by the customer prior to the current booking |
| 20 | reserved\_room\_type | Code of room type reserved. Code is presented instead of designation for anonymity reasons |
| 21 | assigned\_room\_type | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons. |
| 22 | booking\_changes | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation |
| 23 | deposit\_type | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:  No Deposit – no deposit was made;  Non Refund – a deposit was made in the value of the total stay cost;  Refundable – a deposit was made with a value under the total cost of stay. |
| 24 | agent | ID of the travel agency that made the booking |
| 25 | company | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons |
| 26 | day\_in\_waiting\_list | Number of days the booking was in the waiting list before it was confirmed to the customer |
| 27 | customer\_type | Contract - when the booking has an allotment or other type of contract associated to it;  Group – when the booking is associated to a group;  Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;  Transient-party – when the booking is transient, but is associated to at least other transient booking |
| 28 | adr (average daily rate) | average daily rate = Sum of all lodging transaction/Total No of Staying nights |
| 29 | required\_car\_parking\_spaces | Number of car parking spaces required by the customer |
| 30 | total\_of\_special\_requests | Number of special requests made by the customer (e.g. twin bed or high floor) |
| 31 | reservation\_status | Canceled – booking was canceled by the customer;  Check-Out – customer has checked in but already departed;  No-Show – customer did not check-in and did inform the hotel of the reason why |
| 32 | reservation\_status\_date | Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking cancelled or when did the customer checked-out of the hotel |
| 33 | name | Name of the customer |
| 34 | email | Email id of the customer |
| 35 | phone\_number | Phone number of the customer |
| 36 | credit\_card | Credit card number of the customer |

Table 4.1. Dataset Column Information

**4.2 Exploratory Data Analysis**

The ‘is cancelled’ column is to be taken as target so the dataset is to be taken as target. So univariate and multivariate analysis is to be done to find the relation of that target with other columns.

**4.2.1.Univariate Analysis**

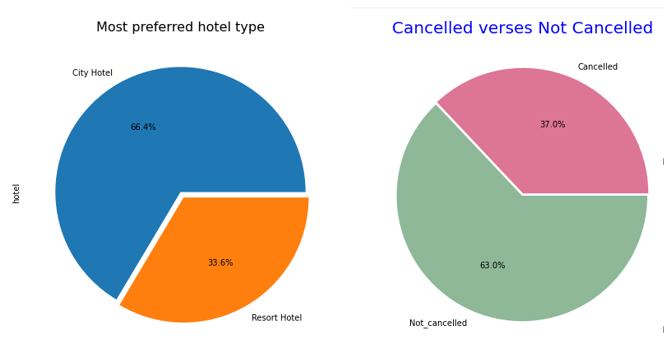
****

Fig.4.2.Pie Chart of Hotel types and Target column

Analysing the dataset , it contains two types of hotels: city hotel and resort hotel. The city hotel contains about 66.4 % data and the resort hotel contains about 33.6% data. So the city hotel is the most preferred hotel type. Also the distribution between cancelled and not cancelled is about 63 to 37 percent. So the data is distributed without any biases w.r.t target column.Cancellations ratio is higher for city hotel than Resort hotel (0.28 to 0.41 percent)

**4.2.2.Bivariate Analysis**

This analysis is done on the given dataset by taking two or more variables.

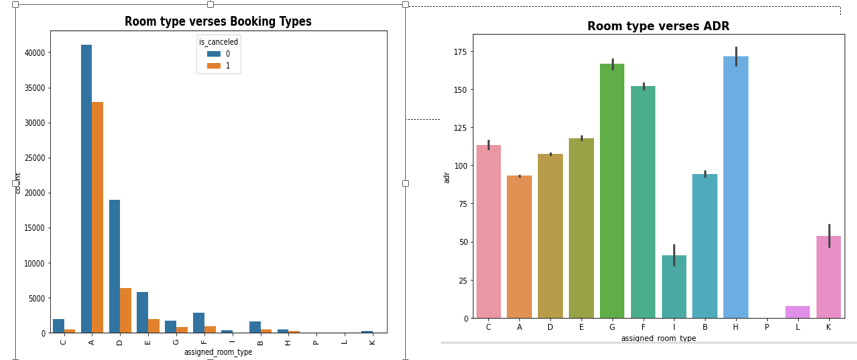


Fig 4.3. Room Type Distributions

The most preferred room type is A. Most demanded room type is A, but better ADR is generated by room type H, G and F also. Hotels should increase the no. of room types G and H to maximise revenue.

On analysing the data set it is found that the number of babies,adults,children, market type, distribution channel all have an impact on cancellation. So these columns will be taken as features in the model building.

After analysing the data, we can see that the cancelled vs non-cancelled bookings ratios are different for the different types of hotels. For the City Hotel, 58.28% of the bookings were non-cancelled vs 41.72% of the bookings that were cancelled, forming a majority. In contrast, for the Resort Hotel, 72.24% of the bookings were non-cancelled whereas 27.76% of the bookings were cancelled. Although, most of the bookings were not cancelled for both hotels; the point to note here is the difference in those numbers for both hotel types

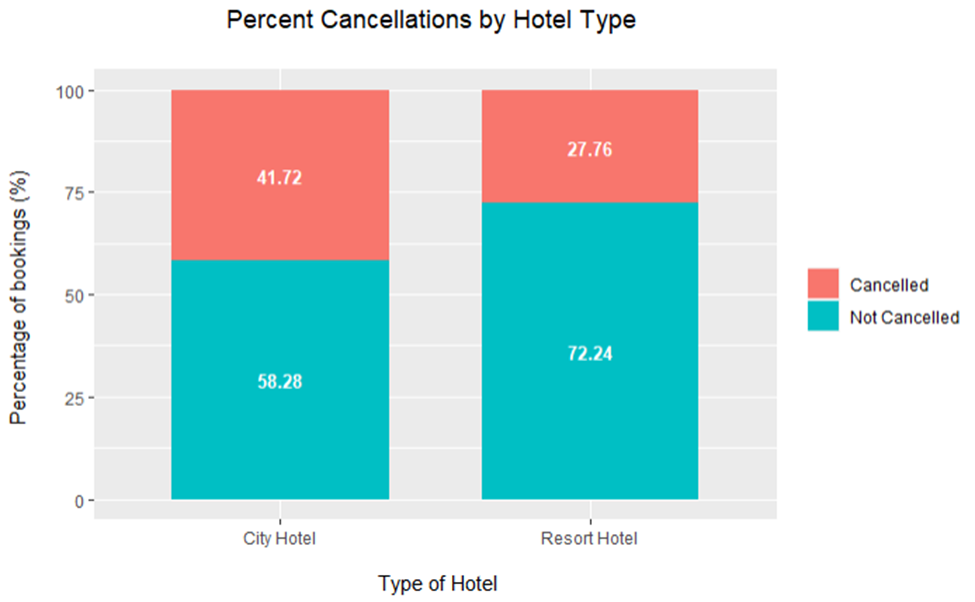


Fig.4.4.Percentage Cancellations Hotel Wise

We are getting a new column called ‘total stay’ by adding stay in week nights column and stay in weekend nights column. The deposit type ‘Non Refund ’and ‘is cancelled ’column are correlated in a counter-intuitive way.

More than double bookings were made in 2016, compared to the previous year. But the bookings decreased by almost 15% the next year.

is\_canceled have two unique values: 1 if booking got cancelled, else 0.Bookings got cancelled 37% of the time. While booking, guests checked-in (did not cancel the booking ) almost 63% of the time.

For Resort hotel, the most popular stay duration is three, two, one, and four days respectively.For City hotel, most popular stay duration is one, two, seven(week), and three respectively

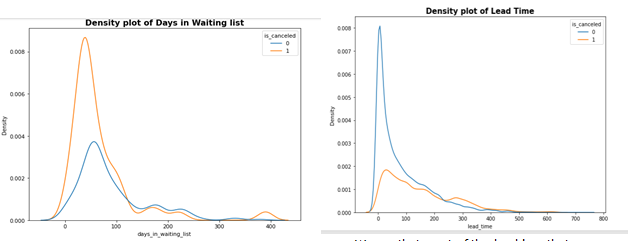


Fig.4.5.Density Plot of Waiting list days and lead time

Most of the bookings that were cancelled had a waiting period of less than 150 days, meanwhile most of bookings that were not cancelled also had a waiting period less than 150 days. Hence this shows that waiting period has no effect on cancellation of bookings. Same applies to Lead time too.

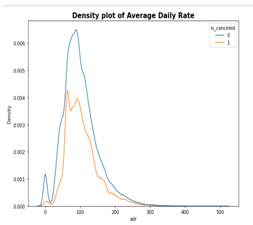


Fig.4.6..Density Plot of ADR

Most of the bookings that are cancelled had an ADR of less than 100, meanwhile most of bookings that are not cancelled also have ADR less than 100. Hence this shows that adr has no effect on cancellation of bookings.

Other inferences from preprocessing are:

* The customer type transient has contributed to the highest number of bookings and most of the bookings were made via Online Travel Agents.
* The customers who made a cancellation in the past have cancelled their bookings more often comparatively.
* Highest number of bookings were made by customers in Portugal, followed by Great Britain, France and Spain.
* More than 60% of the population booked the City hotel
* Most people stay for one, two, or three. More than 60% of guests come under these three options.
* Couple or 2 adults is the most popular accommodation type. So hotels can make arrangement plans accordingly
* The majority of bookings for both hotels had no deposit collected rather than partial or full. The same majority is demonstrated in cancelled bookings for both hotels as well.
* Third-party cancellations are higher than direct hotel bookings, although there is no significant difference in average daily rates.
* Non-repeat guests make most of the cancelled and non-cancelled bookings especially on weekday bookings
* For non-direct bookings, the majority of cancellations on City Hotel, months with higher cancellations generally had a higher average daily rate. The same cannot be proved conclusively for direct bookings

#### Cancellations ratio is higher for city hotel than Resort hotel (0.28 to 0.41 percent)

### City hotel generates high adr when compared to resort hotel

### Maximum arrivals are in the month of April and May

### Maximum arrival is in the year 2017

#### maximum cancellation ratio is for cancellations from 14 to 26

#### Most of the customers come from Portugal, Great Britain, France and Spain

#### Most cancellations are from Transient type

#### Most cancellation ratio is from undefined market segment

#### Most cancellation ratio is from undefined Distribution Channel

### Most cancellation ratio is from Non refund type Deposit

#### P and L type rooms produces maximum cancellations

#### Agent Number 9 has most bookings followed by agent number 240.

**5.Pre-Processing**

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. The various steps followed in hotel booking data pre processing are as follows-

Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

**5.1. Duplicate Value Handling**

The hotel booking dataset has no duplicate rows.

**5.2. Missing Value Handling**

The hotel booking dataset had missing values in the following columns-

* children(4 missing values)

After plotting the distribution of the children column,it is found that the distribution is right skewed.Thus the NaN values were replaced with median value.

* country(488 missing values)

Country column being a categorical variable was replaced with mode value.

* agent(16340 missing values)

Agent column had a right skewed distribution and the missing values were filled with median value.

* company(112593 missing values)

Company column had the largest number of missing values.A significant number of values were missing and filling it may adversely affect the integrity of the dataset.So dropped the company column.

**5.3. Converting Data Types**

The fields, children and agent were of data type float.Since these values cannot be continuous values, these variables were converted from data type float to data type int.

**5.4. Feature Reduction**

Feature reduction is the process of reducing the number of features in a resource heavy computation. Reducing the number of features means the number of variables is reduced making the computer’s work easier and faster.

The following features were dropped as they may not have a significant impact on the model.

* name
* email
* phone\_number
* credit\_card
* arrival\_date\_year
* arrival\_date\_week\_number
* arrival\_date\_day\_of\_month
* arrival\_date\_month
* reservation\_status\_date

**5.5. Feature Extraction**

Inorder to avoid machine learning model suffering from overfitting ,the number of features can be reduced.Feature extraction can also have positive impacts like accuracy improvements,speeding up the training process,improved data visualisation,increased explainability of the model etc.

The following new meaningful features were extracted by making use of some of the existing features-

* Room

The features reserved\_room\_type and assigned\_room\_type were used to determine whether the guests received the same room type which was reserved.This field will contain 1 if the reserved room type is same as the assigned room type and 0 if not.

* net\_cancelled

This column will hold 1 if the guest has cancelled more bookings in the past, when compared to the number of bookings he/she did not cancel and 0 otherwise.This feature utilised previous\_bookings\_not\_canceled and previous\_cancellations columns.

* total\_stay

It is the sum of stays\_in\_weekend\_nights and stays\_in\_week\_nights column.

After extracting these features,the following columns were dropped-

* stays\_in\_weekend\_nights
* stays\_in\_week\_nights
* previous\_bookings\_not\_canceled
* previous\_cancellations
* reserved\_room\_type
* assigned\_room\_type

**5.6. Encoding**

Encoding is the transformation of categorical variables to binary or numerical counterparts.The encoding techniques performed for the hotel booking dataset are -

* Label Encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.The following features were label encoded using LabelEncoder as they had a large number of unique values.

* country
* customer type
* meal
* reservation\_status
* deposit\_type
* market\_segment
* distribution\_channel
* deposit\_type
* One-hot Encoding

One hot encoding creates a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category.

* hotel

Since the field ‘hotel’ has only 2 unique values-Resort Hotel Or City Hotel ,one hot encoding was used for encoding.

**5.7. Outlier handling**

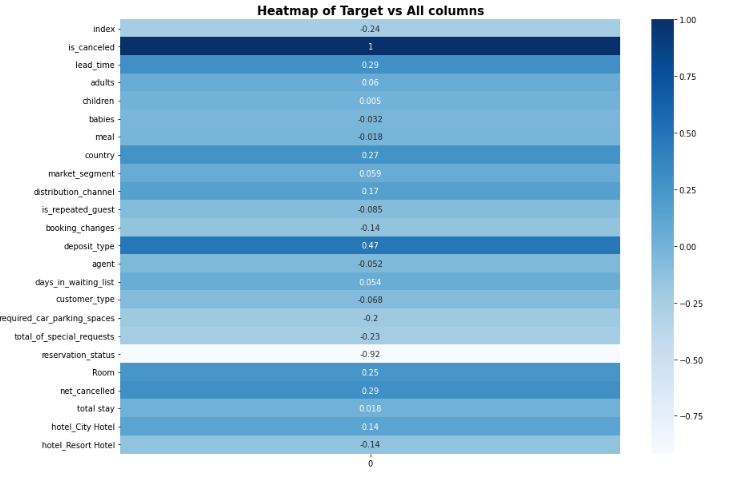
Only continuous numerical columns are taken for outlier handling. There is only 1 continuous numerical column-’adr’(Average daily rate). So Outliers are present only in one of the columns ‘adr’. There are two outliers present in this column. A boxplot of the adr column is drawn to find the outliers. The value 5400 is an outlier and is far higher than the rest of the values so it is dropped. Also dropped values less than 0 keeping only rows less than 5400 and greater than 0 in the dataset. Finally the shape of the dataset becomes (119388, 24) after all the pre-processing steps.

**5.8. Scaling on the Dataset**

Scaling of the dataset is only one ‘adr’ (average daily rate) column. It is the only column with continuous values in the dataset. Standard scaling is done on the ‘adr’ column. There is only 1 continuous column (ADR) in the dataset, so Principal component Analysis (PCA) is not possible in the dataset.

**5.9.Heat Map Analysis**

Heat map is plotted to analyse the correlation of the target column with all the feature columns. The heat map values are from -0.92 to 0.47. The value -0.92 is highly negatively correlated ie. ‘Reservation status’ column . So that column is also dropped from the dataset. Also the country column is dropped from the dataset as it contains a lot of unique values in the dataset. The shape of the dataset becomes (119388, 22). So this preprocessed dataset is used for model building

Fig.5.1. Heat Map of target column vs features

**6.Model Building**

Total of 9 models were tried out in the preprocessed dataset. They are

***1.Decision Tree Model***

***2.Random Forest***

***3.Logistic Regression***

***4.KNN***

***5.Gradient Boosting***

***6.XGB Classifier***

***7.LGBM Classifier***

***8.Extra Trees Classifier***

***9.Adaboost Classifier***

The Accuracy score, precision, Recall, And F1 score of the models are found out of these models and they are analysed to find the best model. Also the confusion matrix is found out. The confusion matrix is analysed to find the accuracy score of the model.

* Accuracy = (True Positive + True Negative) / (True Positive + False Negative + False Positive + True Negative)
* Precision=True positives/(True Positive+False Positive)
* Recall=True positives/(True Positive+False Negatives)
* The F1 Score is the 2\*((precision\*recall)/(precision+recall)

The model is built with all the 21 features in ‘X’ and the target ‘is cancelled’ as the target column. Test size is taken as 0.25 and random state is fixed as 42

**6.1. Model 1-Decision Tree Algorithm**

Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

For our dataset the results are obtained as follows in the Decision Tree model.

Accuracy Score of Decision Tree is : 0.8294971018862868

***Confusion Matrix*** : [[16154 2545]

[ 2544 8604]]

***Classification Report***:

precision recall f1-score support

0 0.86 0.86 0.86 18699

1 0.77 0.77 0.77 11148

accuracy 0.83 29847

macro avg 0.82 0.82 0.82 29847

weighted avg 0.83 0.83 0.83 29847

**6.2. Model 2-Random Forest algorithm**

Next model was built using random forest algorithm. It is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

In our model accuracy score of random forest is : 0.8637383991690957

***Confusion Matrix*** : [[17157 1542]

[ 2547 8601]]

***Classification Report :***

precision recall f1-score support

0 0.87 0.92 0.89 18699

1 0.85 0.77 0.81 11148

accuracy 0.86 29847

macro avg 0.86 0.85 0.85 29847

weighted avg 0.86 0.86 0.86 29847

**6.3. Model 3.-Logistic Regression**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

In our model accuracy score of logistic regression is obtained as : 0.7700606426106477

***Confusion Matrix*** : [[17145 1554]

[ 5309 5839]]

***Classification Report :***

precision recall f1-score support

0 0.76 0.92 0.83 18699

1 0.79 0.52 0.63 11148

accuracy 0.77 29847

macro avg 0.78 0.72 0.73 29847

weighted avg 0.77 0.77 0.76 29847

**6.4. Model 4.-KNN Algorithm**

A model was built using K-nearest neighbors (KNN) algorithm.It is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification of predictive problems in industry.

In our model accuracy score of KNN is : 0.8054745870606761

***Confusion Matrix*** : [[16090 2609]

[ 3197 7951]]

***Classification Report*** :

precision recall f1-score support

0 0.83 0.86 0.85 18699

1 0.75 0.71 0.73 11148

accuracy 0.81 29847

macro avg 0.79 0.79 0.79 29847

weighted avg 0.80 0.81 0.80 29847

**6.5.Model 5.-Gradient Boosting Algorithm**

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. There are mainly two types of error, bias error and variance error. Gradient boost algorithm help us minimise bias error of the model.Gradient boosting has a fixed base estimator i.e., Decision Trees whereas in AdaBoost we can change the base estimator according to our needs.A gradient boosting classifier is used when the target column is binary.

In our model accuracy score of gradient boosting classifier is : 0.8252755720842966

***Confusion Matrix :***

[[17525 1174]

[ 4041 7107]]

***Classification Report :***

precision recall f1-score support

0 0.81 0.94 0.87 18699

1 0.86 0.64 0.73 11148

accuracy 0.83 29847

macro avg 0.84 0.79 0.80 29847

weighted avg 0.83 0.83 0.82 29847

**6.6. Model 6-XGB Classifier**

XGBoost (Extreme Gradient Boosting Algorithm) is a powerful and popular machine learning algorithm that is often used to win machine learning competitions. It is an implementation of the gradient boosting algorithm that is designed to be highly efficient and scalable. This involves dividing your data into training and testing sets, building a model on the training set, and then evaluating the model on the testing set. Another popular method is to use a hold-out set that is not used for training the model. The model is then evaluated on this hold-out set.

In our model accuracy score of XGB classifier is : 0.8434348510738098

***Confusion Matrix :***

[[17479 1220]

[ 3453 7695]]

***Classification Report :***

precision recall f1-score support

0 0.84 0.93 0.88 18699

1 0.86 0.69 0.77 11148

accuracy 0.84 29847

macro avg 0.85 0.81 0.82 29847

weighted avg 0.85 0.84 0.84 29847

**6.7.Model 7-LGBM Classifier**

Light GBM (Gradient Boosting Machine) is a gradient boosting framework that uses tree based learning algorithms. It is designed with the following advantages: Faster training speed and higher efficiency, Lower memory usage, Better accuracy, Support of parallel and GPU learning, Capable of handling large-scale data. Light GBM can handle the large size of data and takes lower memory to run. Another reason why Light GBM is so popular is because it focuses on accuracy of results.

In our model accuracy score of LGBM classifier is : 0.8361979428418267

***Confusion Matrix :***

[[16784 1915]

[ 2974 8174]]

***Classification Report :***

precision recall f1-score support

0 0.85 0.90 0.87 18699

1 0.81 0.73 0.77 11148

accuracy 0.84 29847

macro avg 0.83 0.82 0.82 29847

weighted avg 0.83 0.84 0.83 29847

**6.8. Model 8-Extra Tree Classifier**

Second best model was built using extra tree classifier algorithm. It is like the random forest’s algorithm, which creates many decision trees, but the sampling for each tree is random, without replacement. This creates a dataset for each tree with unique samples. A specific number of features, from the total set of features, are also selected randomly for each tree. The most important and unique characteristic of extra trees is the random selection of a splitting value for a feature.

In our model accuracy score of extra trees classifier is : 0.8596843903909941

***Confusion Matrix*** : [[17074 1625]

[ 2563 8585]]

***Classification Report :***

precision recall f1-score support

0 0.87 0.91 0.89 18699

1 0.84 0.77 0.80 11148

accuracy 0.86 29847

macro avg 0.86 0.84 0.85 29847

weighted avg 0.86 0.86 0.86 29847

**6.9. Model 9-AdaBoost Classifier**

Then a model was created using the AdaBoost classifier algorithm. It is also called Adaptive Boosting. It is a technique in Machine Learning which is used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means Decision tree with only 1 split. These trees are also called Decision Stumps.

In our model accuracy score of adaBoost classifier is : 0.8456796327939157

***Confusion Matrix :*** [[16780 1919]

[ 2687 8461]]

***Classification Report :***

precision recall f1-score support

0 0.86 0.90 0.88 18699

1 0.82 0.76 0.79 11148

accuracy 0.85 29847

macro avg 0.84 0.83 0.83 29847

weighted avg 0.84 0.85 0.84 29847

The scores of all 9 models are analysed and then the best model is found out.

| **Model** | **Accuracy Scores** |
| --- | --- |
| Random Forest Classifier | 0.863 |
| Extra Trees Classifier | 0.8597 |
| Ada Boost Classifier | 0.8484 |
| LGBM | 0.844339 |
| XgBoost | 0.843435 |
| Decision Tree Classifier | 0.828861 |
| Gradient Boosting Classifier | 0.825276 |
| KNN | 0.805475 |
| Logistic Regression | 0.770061 |

Table 6.1.Accuracy scores of Models

From the comparison of accuracy scores it is found out that Random Forest Algorithm gives the best Accuracy score. So the Random Forest algorithm model is selected as the best model.

**6.10. Cross Validation**

Cross-validation is a statistical method used to estimate the performance of machine learning models. It is a method for assessing how the results of a statistical analysis will generalise to an independent data set.Cross Validation of the models is done using k-fold validator and stratified k fold validator

In the technique of K-Fold cross-validation, the whole dataset is partitioned into K parts of equal size. Each partition is called a “Fold“.So as we have K parts we call them K-Folds. One Fold is used as a validation set and the remaining K-1 folds are used as the training set.The technique is repeated K times until each fold is used as a validation set and the remaining folds as the training set.

Stratified K-Fold is an enhanced version of K-Fold cross-validation which is mainly used for imbalanced datasets. Just like K-fold, the whole dataset is divided into K-folds of equal size.But in this technique, each fold will have the same ratio of instances of target variable as in the whole datasets.

The Cv scores are obtained between 0.75 to 0.59 for different models using k-fold validator.After taking 10 important features the Cv scores are obtained between 0.79 to 0.61 for different models using k-fold validator. After feature selection the cross validation of models is again done to verify the accuracy of the models. The CV scores are varying between 0.71 to 0.61 in k-fold and 0.74 to 0.63 using stratified k fold after 10 feature selection.

´Random forest is giving better results compared to other models as per cross validation

**6.11. Feature Selection**

Feature selection is the process of identifying and selecting a subset of input features that are most relevant to the target variable.There are two popular feature selection techniques that can be used for numerical input data and a categorical target variable.They are: ANOVA-f Statistic and Mutual Information Statistics.

F-test, is a class of statistical tests that calculate the ratio between variances values, such as the variance from two different samples. ANOVA is used when one variable is numeric and one is categorical, such as numerical input variables and a classification target variable in a classification task.Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable

**6.11.1.Select-K Best Feature Selection Library**

Scikit-learn API provides Select K Best class for extracting best features of given dataset. The SelectKBest method selects the features according to the k highest score. By changing the 'score\_func' parameter we can apply the method for both classification and regression data. Selecting the best features is an important process when we prepare a large dataset for training. It helps us to eliminate less important parts of the data and reduce training time.The scikit-learn machine library provides an implementation of the ANOVA f-test in the f\_classif() function. This function can be used in a feature selection strategy, such as selecting the top k most relevant features (largest values) via the Select K Best class. In the dataset we can define the Select K Best class to use the f\_classif() function and select all features, then transform the train and test sets.Then print the scores for each variable (larger is better) and plot the scores for each variable as a bar graph to get an idea of how many features we should select.

| **Parameters** | ***score\_fun:callable, default=f\_classif***  Function taking two arrays X and y, and returning a pair of arrays (scores, pvalues) or a single array with scores. Default is f\_classif (see below “See Also”). The default function only works with classification tasks. |
| --- | --- |
| ***kint or “all”, default=10***  Number of top features to select. The “all” option bypasses selection, for use in a parameter search. |
| ***Attributes*** | ***scores\_array-like of shape (n\_features,)***  Scores of features. |
| ***pvalues\_array-like of shape (n\_features,)***  p-values of feature scores, None if score\_func returned only scores. |
| ***n\_features\_in\_int***  Number of features seen during fit. |
| ***feature\_names\_in\_ndarray of shape (n\_features\_in\_,)***  Names of features seen during fit. Defined only when X has feature names that are all strings. |

Table 6.2.Feature Selection Parameters

The feature selection scores (Fs scores) of all Features in the dataset is found out. The features are arranged in descending order as shown in the figure.The best 10 features are found out as

1. lead\_time
2. distribution\_channel
3. booking\_changes
4. deposit\_type
5. required\_car\_parking\_spaces
6. total\_of\_special\_requests
7. Room
8. net\_cancelled
9. hotel\_City Hotel
10. hotel\_Resort Hotel

The 9 models are again evaluated with 10 features only and accuracy scores are obtained

| **Model** | **Accuracy Scores with 10 Features** |
| --- | --- |
| Random Forest Classifier | 0.787483 |
| Extra Trees Classifier | 0.786880 |
| Ada Boost Classifier | 0.786846 |
| LGBM | 0.786813 |
| XgBoost | 0.786310 |
| Decision Tree Classifier | 0.779241 |
| Gradient Boosting Classifier | 0.774751 |
| KNN | 0.766878 |
| Logistic Regression | 0.765404 |

Table 6.3.Accuracy scores after feature selection

The accuracy scores are in the range of 0.78 to 0.76 with 10 features. So these 10 features can be used for Web Deployment and Booking prediction

**6.12. Hyper parametric Tuning of the models**

Hyper parameter tuning refers to adjusting the settings of an algorithm(hyper parameters) to optimise the performance. Using Scikit-Learn’s Randomised Search CV method, we can define a grid of hyper parameter ranges, and randomly sample from the grid, performing K-Fold CV with each combination of values.

The parameters used are

n\_estimators : ***int, default=100:*** This is perhaps the most important parameter. This represents the number of trees you want to build within a random forest before calculating the predictions. Usually, the higher the number, the better, but this is more computationally expensive.

max\_features : ***{“auto,” “sqrt,” “log2”}, int or float, default=” auto”:*** This represents the number of features that are considered on a pre-split level when finding the best split. This improves the model's performance as each tree node is now considering a higher number of options.

max\_depth :***int, default=None*** This is used to select how deep you want to make each tree in the forest. The deeper the tree, the more splits it has, and it captures more information about the data.

min\_samples\_split :***int or float, default=2:*** This specifies the minimum number of samples that must be present from your data for a split to occur.

min\_samples\_leaf : ***int or float, default=1:*** This parameter helps determine the minimum required number of observations at the end of each decision tree node in the random forest to split it.

bootstrap : method for sampling data points (with or without replacement)

Fitting 5 folds for each of 100 candidates, totalling 500 fits

The following variables are tried in the model for hyperparameter tuning.

param\_distributions='bootstrap': [True, False],

'max\_depth': [10, 20, 30, 40, 50, 60,70, 80, 90, 100, 110,120]

'max\_features': ['auto', 'sqrt'],

'min\_samples\_leaf': [1, 3, 4],

'min\_samples\_split': [2, 6, 10],

'n\_estimators': [5, 20, 50, 100]

Best Parameters are obtained as

1. 'n\_estimators': 100,
2. 'min\_samples\_split': 10, ‘
3. min\_samples\_leaf': 4,
4. 'max\_features': 'sqrt',
5. 'max\_depth': 20,
6. 'bootstrap': True

To determine if random search yielded a better model, compared the base model with the best random search model.

Percentage Increase of Accuracy Score of Random Forest before and after Hyper Parametric Tuning : 0.63

There was an improvement of nearly 1% in the accuracy score of the Random Forest Model after performing hyper parameter tuning.

**7.Web Deployment**

The web deployment is done after selecting the best model ‘Random forest Algorithm’ and doing hyperparameter tuning on the model. The Dataset with only 10 feature columns and target columns is saved and is used for web deployment.The 10 features which are used for web deployment are

10 Best features are

1. hotel City Hotel
2. lead time
3. distribution channel
4. booking changes
5. deposit type
6. required ca parking spaces
7. total\_of special requests
8. Room
9. net cancelled
10. hotel Resort Hotel

There are both numerical and categorical columns which are to be entered from the user. The numerical columns are

1. Lead time
2. Required Car parking spaces
3. Total number of special requests
4. Number of times booking changes

The categorical columns are

1. Distribution channel
2. Deposit type

These columns are label encoded in preprocessing. So they need to be entered in its original form and then converted as encoded values into the model.

The Hotel column is one hot encoded in preprocessing and thus two columns hotel City Hotel and hotel Resort Hotel are created. So in the user interface page there is only one column. Hotel type and from that input data is taken back to original form

if hotel=="City\_Hotel":

hotel\_City\_Hotel=1; hotel\_Resort\_Hotel =0

elif hotel=="Resort\_Hotel":

hotel\_City\_Hotel=0 ; hotel\_Resort\_Hotel =1

The Room column is a preprocessed column so it is converted back to its original form.The value for Room is obtained from assigned room vs reserved room comparison

If Assign\_type==Reserved\_type:

Room=1

else: Room=0

The Net cancelled columns are splitted into their original columns. Net canceled is obtained from comparison of previous bookings and previous bookings not cancelled values if

if previous\_cancell>previous\_cancell\_not:

net\_cancelled=1

else: net\_cancelled=0

For web deployment App file , Model File , html file and css style sheets are created. The best Model i.e. Random forest model is given to the model file and a pickle file is created. In the App file the features are taken as inputs and then some of the features are encoded and given to the model. The website appearance is styled in index1.html and index2.html and css files to male the website more attractive. After the successful local hoisting of the website it is globally hosted using python anywhere website. Html file pickle file and app file are added to the site css File is added under static subfolder. Some background texture is also added to make the website more attractive.Then website is successfully hoisted under the domain name <https://dsaictbatch5.pythonanywhere.com>. Some Random values are given for the features columns and the website classify the result as ‘Booking may be cancelled’ or ‘Booking may not be cancelled’

**8. Result**

The booking cancellation forecasting of hotel booking dataset using a machine learning model ,was implemented using Random Forest Classifier. Various algorithms like Extra Trees Classifier,Ada boost Classifier,Xgboost,LGBM,Decision Tree Classifier,Gradient Boosting Classifier,KNN,Logistic Regression and Random Forest Classifier were used ,out of which Random Forest Classifier gave the highest accuracy value of 0.863.

| **Model** | **Accuracy Scores** |
| --- | --- |
| Random Forest Classifier | 0.863 |
| Extra Trees Classifier | 0.8597 |
| Ada Boost Classifier | 0.848427 |
| LGBM | 0.844339 |
| XgBoost | 0.843435 |
| Decision Tree Classifier | 0.8281 |
| Gradient Boosting Classifier | 0.825276 |
| KNN | 0.805475 |
| Logistic Regression | 0.770061 |

Table 8.1.Accuracy scores of Models

After performing K Fold cross validation on the created models,the K Fold Cross validation scores of the models were found to be the following.

| **Model** | **Kfold Cv scores** |
| --- | --- |
| Random Forest Classifier | 0.754105 |
| XgBoost | 0.736037 |
| LGBM | 0.679239 |
| Gradient Boosting Classifier | 0.677304 |
| Extra Trees Classifier | 0.665041 |
| Logistic Regression | 0.664322 |
| Ada Boost Classifier | 0.653348 |
| Decision Tree Classifier | 0.641286 |
| KNN | 0.570502 |

Table 8.2.K fold Cv scores of Models

Cross validation using Stratified K Fold was also performed for all the models and the cross validation scores are as follows.

| **Model** | **Stratified Kfold Cv scores** |
| --- | --- |
| Random Forest Classifier | 0.797057 |
| XgBoost | 0.778252 |
| LGBM | 0.742872 |
| Logistic Regression | 0.738366 |
| Gradient Boosting Classifier | 0.714535 |
| Extra Trees Classifier | 0.707357 |
| Ada Boost Classifier | 0.684775 |
| Decision Tree Classifier | 0.668727 |
| KNN | 0.607515 |

Table 8.3.Stratified K fold Cv scores of Models

For finding the most meaningful features,feature selection was performed using SelectKBest and 10 features with the top 10 feature scores were selected from a total of 21 features.

The following are the 10 features obtained after considering the feature selection scores.

1. deposit\_type
2. lead\_time
3. net\_cancelled
4. Room
5. total\_number\_of\_special\_requests
6. required\_car\_parking\_spaces
7. distribution\_channel
8. booking\_changes
9. hotel\_resort\_hotel
10. hotel\_city\_hotel

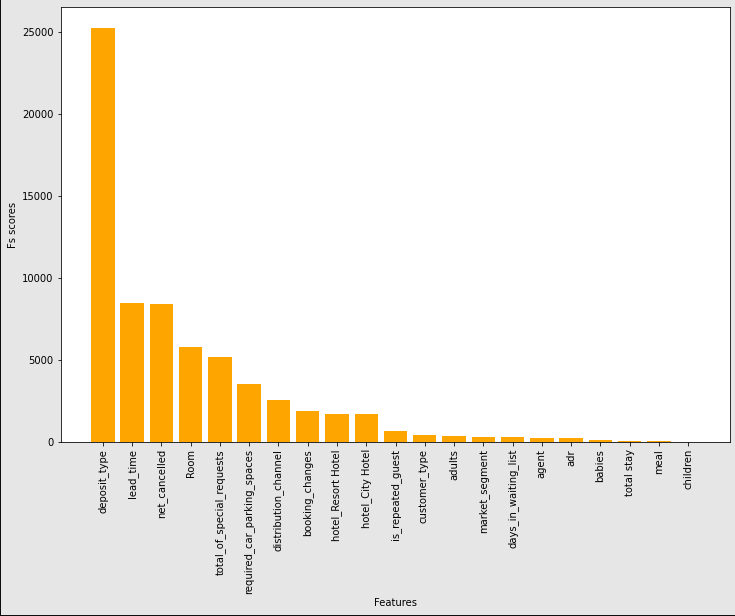


Fig.8.1.Fs scores of all features in descending order

After building the model with the above 10 features,the accuracy scores for all the models were as follows.

| **Model** | **Accuracy Scores with 10 Features** |
| --- | --- |
| Random Forest Classifier | 0.787483 |
| Extra Trees Classifier | 0.786880 |
| Ada Boost Classifier | 0.786846 |
| LGBM | 0.786813 |
| XgBoost | 0.786310 |
| Decision Tree Classifier | 0.779241 |
| Gradient Boosting Classifier | 0.774751 |
| KNN | 0.766878 |
| Logistic Regression | 0.765404 |

Table 8.4..Accuracy scores after feature selection

The cross validation scores of all models are found with 10 features only using both k fold and stratified K fold validation. The Cv scores are obtained as follows

| **Model** | **Kfold CV scores-10 Features** |
| --- | --- |
| Random Forest Classifier | 0.710426 |
| Gradient Boosting Classifier | 0.691018 |
| Xgboost | 0.671343 |
| Ada Boost Classifier | 0.663552 |
| Decision Tree Classifier | 0.663150 |
| Extra Trees Classifier | 0.660637 |
| Logistic Regression | 0.660612 |
| LGBM | 0.657119 |
| KNN | 0.609591 |

Table 8.5.K fold CV scores of Models-10 Features

Random Forest Classifier has got the highest K fold CV score of 0.710426.

| **Model** | **Stratified Kfold CV scores-10 Features** |
| --- | --- |
| Random Forest Classifier | 0.749716 |
| Logistic Regression | 0.749063 |
| XgBoost | 0.730635 |
| LGBM | 0.721120 |
| Gradient Boosting Classifier | 0.707352 |
| AdaBoost Classifier | 0.706805 |
| Extra Trees Classifier | 0.703564 |
| Decision Tree Classifier | 0.668727 |
| KNN | 0.631655 |

Table 8.6.Stratified K fold CV scores of Models-10 Features

Random Forest Classifier has got the highest Stratified K fold CV score of 0.749716.

Hyper parameter tuning was done in the Random Forest Model with 10 features using Random Forest Regressor and RandomizedSearchCV.The accuracy scores before and after the hyper parameter tuning are as follows.

Best Parameters obtained are **n\_estimators= 100, min\_samples\_split= 10, min\_samples\_leaf=4, max\_features=sqrt, max\_depth=20, bootstrap= True**

Accuracy Score of Random Forest before Tuning : 0.7849700137367239

Accuracy Score of Random Forest After Tuning : 0.7913023084397092

Percentage Increase of Accuracy Score of Random Forest before and after Hyper Parametric Tuning : 0.63

There was an improvement of nearly 1% in the accuracy score of the Random Forest Model after performing hyper parameter tuning.

**9. Conclusion**

The Hotel booking dataset comprises 36 variables and 119390 observations.It includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, etc. Analysing the dataset , it contains two types of hotels namely, city hotel and resort hotel. The city hotel contains about 66.4 % data and the resort hotel contains about 33.6% data. is\_canceled column can be taken as a target column for classification model. The distribution between cancelled and not cancelled is about 63 to 37 percent. It was found that the customers who made a cancellation in the past have cancelled their bookings more often comparatively. Highest number of bookings were made by customers in Portugal, followed by Great Britain, France and Spain.Missing values were handled in agent, children and country column. Exploratory data analysis was done using this dataset to find out various relationships.Encoding of the dataset was done using label encoding and one hot encoding. Some additional columns are created as a part of feature engineering and some unnecessary columns are also dropped.

Model building was done using different models such as Decision Tree, Random Forest, Logistic Regression, KNN, Gradient Boosting, XGBClassifier, LGBM Classifier,Extra Trees Classifier and Adaboost Classifier out of this Random forest was the one which gave the maximum accuracy score this may be because random forests efficiency is particularly notable in large data sets,unlike other models Random forest does not overfit with more features. Feature selection was done on the dataset and also the 10 best features are found out. The Web deployment was done using selected features. Select-K\_best in the sklearn library helps to achieve it. The 10 features that contribute very much to the model as mostly related to customer preferences. So in future bookings in these hotels special consideration should be given to customer requirements so that unnecessary cancellations can be avoided. Cross validation and hyperparameter tuning is also done on the models to improve the efficiency of the models. In Cross validation Random Forest is giving better results than other models and the hyperparameter tuning slightly increased the model efficiency.

Some other types of Classification Algorithm like Naïve Bayes,Stochastic Gradient Descent, SVM linear, SVM RBF are not tried on the dataset. Trying these classification Algorithms and calculating Accuracy scores and trying to improve these models using hyperparameter tuning will improve the model efficiency. These can be done in future studies. Also Collecting datasets from other service industries and building models on them will help to reduce the cancellation effects on these industries. The Booking cancellation Prediction can also be tried by other E-commerce platforms, so they can reduce unnecessary transportation and packing charges.

**References**

[1] Application of Machine Learning in the Hotel Industry: A Critical Review, Dr. Eid Tourism, Archaeology Department, College of Arts, University of Hail, P.O.Box 2440,Hail, Saudi Arabia, 2020.

[2] Hotel Booking Prediction using Machine Learning,Pranav Kumar, Sarthak Sharma,International Journal for Research in Applied Science & Engineering Technology (IJRASET),2022

[3] Predicting Hotel Bookings Cancellation with machine learning classification model, Nuno Antonio, Ann de Almeida, Luis Nunes, IEEE International Conference on Machine Learning and Application, 2017.

[4] An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations , Nuno Antonio , Ana de Almeida, Luis Nunes,Data Science Journal, 18(1), p.32, [2019]

[5] Predicting hotel booking cancellations to decrease uncertainty and increase revenue,Antonio, N., de Almeida, A. and Nunes, L,Tourism & Management Studies, Vol. 13, No. 2, pp. 25-39,[2017]

[6] Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis,Soo Y. Kim,The Service Industries Journal Vol. 31, No. 3, 44 – 46, [2011]

[7] Using Data Science to Predict Hotel Booking Cancellations,Nuno Antonio , Ana de Almeida, Luis Nunes Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry pp.140-166,[2016]

[8] Prediction of Hotel Booking Cancellation using CRISP-DM Zharfan Akbar Andriawan *et al*,4th International Conference on Informatics and Computational Sciences (ICICoS) [2020]

[9] Prediction of hotel booking cancellation using deep neural network and logistic regression algorithm,Nugroho Adi Putro1, Rendi Septian Widiastuti, Mawadatul Maulidah, Hilman Ferdinandus Pardede, jurnal Techno nusa mandiri Vol.18,[2021]

[10] Predictive models for hotel booking cancellation: a semi-automated analysis of the literature,Nuno António,Ana de Almeida, Luis Nunes,Tourism & Management Studies, 15(1), 7-21[2019]