**IST-687: Introduction to Data Science**

**HOTEL STAY CANCELLATION PREDICTION MODELLING**



**Date: 9th December 2021**

**Project by:**

Hemang Gala

Soham Nanavati

Ryan Song

Nidhee Patil

Akash Malde

Contents

[INTRODUCTION AND SCOPE OF THE PROJECT 2](#_Toc89980220)

[PROBLEM STATEMENT 2](#_Toc89980221)

[PRIMARY GOAL OF THE PROJECT 3](#_Toc89980222)

[BUSINESS QUESTIONS 3](#_Toc89980223)

[THE DATASET (Descriptive statistics) 3](#_Toc89980224)

[DATA ACQUISITION, CLEANING AND TRANSFORMATION 8](#_Toc89980225)

[EXPLORATORY DATA ANALYSIS 10](#_Toc89980226)

[CLASSIFICATION AND REGRESSION TREE 19](#_Toc89980227)

[ASSOCIATION RULE MINING 25](#_Toc89980228)

[SVM: Support Vector Machine 29](#_Toc89980229)

[ACTIONABLE INSIGHTS 34](#_Toc89980230)

# INTRODUCTION AND SCOPE OF THE PROJECT

This project deals with the cancellation of bookings at a hotel. The cancellation of bookings at a hotel greatly impacts the financial forecasting and affect the revenue as well. The project utilizes the real-life dataset acquired to perform data science tasks like data exploration, data cleaning, transformation and munging, data visualization and data modelling to gain insights on the cancellation of bookings and the hotel and find out if the future cancellations can be predicted. Several R libraries have been used in this project for the abovementioned tasks. The project aims at finding out the trends for booking cancellations using exploratory data analysis and rule modelling and predicting future booking cancellations by developing machine learning models to predict the same.

# PROBLEM STATEMENT

With the increase in cancellations of bookings at the hotel, there is a need to find out the trends behind these cancellations. Moreover, in order to be ready for cancellations in the future, there is a need for a sophisticated model which can be utilized to detect any booking cancellations in the future and prepare the hotel management to be better prepared to tackle such booking cancellations.

# PRIMARY GOAL OF THE PROJECT

The main goal of the project is to find out the trends (if any) behind the cancellation of hotel stay bookings. Also, there is a need to find out the features of the bookings and the people booking which lead to cancellations and document the same. Data Visuals will be key to finding these trends and providing them will be necessary. Furthermore, highly accurate machine learning models will be created to better predict the cancellation of bookings.

# BUSINESS QUESTIONS

The following business questions are addressed in this project:

1. Which features of the booking are more frequently appearing in cancelled bookings?
2. How do different features impact the cancellation of bookings?
3. Which machine learning model has the highest prediction accuracy and should be used by the hotel for the prediction of booking cancellations?

# THE DATASET (Descriptive statistics)

Each row in the dataset: Hotel booking

Explanation: each row consists of the values of different variables that comes into consideration while booking a hotel and if the bookings were cancelled. This data consists of data from various countries.

The data consists of 40060 rows and 20 columns. The data consists of seven categorical variables and rest numeric variables. The description of each variable is as follows:

• **IsCanceled**: Categorical Value indicating if the booking was canceled (1) or not (0)

• **LeadTime**: Integer, Number of days that elapsed between the entering date of the booking into and the arrival date

Min.: 0.00

1st Qu.: 10.00

Median: 57.00

Mean : 92.68

3rd Qu.:155.00

Max. :737.00

• **StaysInWeekendNights**: Integer, Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

Min. : 0.00

1st Qu.: 0.00

Median : 1.00

Mean : 1.19

3rd Qu.: 2.00

Max. :19.00

• **StaysInWeekNights**: Integer, Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel

Min. : 0.000

1st Qu.: 1.000

Median : 3.000

Mean : 3.129

3rd Qu.: 5.000

Max. :50.000

• **Adults**: Integer, Number of adults

Min. : 0.000

1st Qu.: 2.000

Median : 2.000

Mean : 1.867

3rd Qu.: 2.000

Max. :55.000

• **Children**: Integer, Number of children

Min. : 0.0000

1st Qu.: 0.0000

Median : 0.0000

Mean : 0.1287

3rd Qu.: 0.0000

Max. :10.0000

• **Babies**: Integer, Number of babies

Min. :0.0000

1st Qu.:0.0000

Median :0.0000

Mean :0.0139

3rd Qu.:0.0000

Max. :2.0000

• **Meal**: Categorical, Type of meal booked. Categories are presented in standard hospitality meal packages: 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: Categorical, Country of origin. Categories are represented in the ISO 3155– 3:2013 format

• **MarketSegment**: Categorical, Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

• **IsRepeatedGuest**: Categorical, Value indicating if the booking name was from a repeated guest (1) or not (0)

• **PreviousCancellations**: Integer, Number of previous bookings that were cancelled by the customer prior to the current booking

Min.: 0.0000

1st Qu.: 0.0000

Median: 0.0000

Mean: 0.1017

3rd Qu.: 0.0000

Max. :26.0000

• **PreviousBookingsNotCanceled**: Integer, Number of previous bookings not cancelled by the customer prior to the current booking

Min. : 0.0000

1st Qu.: 0.0000

Median : 0.0000

Mean : 0.1465

3rd Qu.: 0.0000

Max. :30.0000

• **ReservedRoomType**: Categorical, Code of room type reserved. Code is presented instead of designation for anonymity reasons

• **AssignedRoomType**: Categorical, 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

• **BookingChanges**: Integer, 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

• **DepositType**: Categorical, 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.

• **CustomerType**: Categorical, Type of booking, assuming one of four categories: 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

• **RequiredCardParkingSpaces**: Integer, Number of car parking spaces required by the customer

Min. :0.0000

1st Qu.:0.0000

Median :0.0000

Mean :0.1381

3rd Qu.:0.0000

Max. :8.0000

• **TotalOfSpecialRequests**: Integer, Number of special requests made by the customer (e.g. twin bed or high floor)

Min. :0.0000

1st Qu.:0.0000

Median :0.0000

Mean :0.6198

3rd Qu.:1.0000

Max. :5.0000

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Minimum | 1st Quartile | Median | Mean | 2nd Quartile | Maximum |
| LeadTime | 0.0 | 10.0 | 57 | 92.68 | 155 | 937 |
| StaysInWeekendNights | 0.0 | 0.0 | 1.0 | 1.19 | 2.0 | 19.0 |
| StaysInWeekNights | 0.0 | 1.0 | 3.0 | 3.129 | 5.0 | 50.0 |
| Adults | 0.0 | 2.0 | 2.0 | 1.867 | 2.0 | 55.0 |
| Children | 0.0 | 0.0 | 0.0 | 0.1287 | 0.0 | 10.0 |
| Babies | 0.0 | 0.0 | 0.0 | 0.0139 | 0.0 | 2.00 |
| PreviousCancellations | 0.0 | 0.0 | 0.0 | 0.1017 | 0.0 | 26.0 |
| PreviousBookingsNotCanceled | 0.0 | 0.0 | 0.0 | 0.1465 | 0.0 | 30.0 |
| RequiredCarParkingSpaces | 0.0 | 0.0 | 0.0 | 0.1381 | 0.0 | 8.0 |
| TotalOfSpecialRequests | 0.0 | 0.0 | 0.0 | 0.6198 | 1.0 | 5.0 |
|  |  |  |  |  |  |  |

# DATA ACQUISITION, CLEANING AND TRANSFORMATION

**Data Acquisition:**

We had been given the dataset in form of a link. This data set consists of real-life hotel stay data.

After the initial exploration of the given data, we have imported the dataset using the read\_csv() function to convert it into a data frame in R.

A picture containing shape

Description automatically generated

Snippet of the data:

Application, table

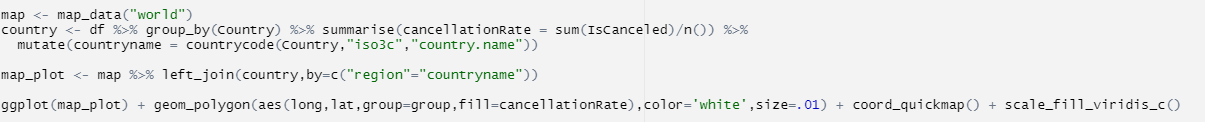
Description automatically generated

**Data Cleaning:**

While creating visualizations for the dataset, we have considered mutating some of the similar variables such as StayinWeekendnights and StayInWeekDaynights, to get the total number of days the guests are staying at the resort. We even combined the variables Adults, Children and Babies to get total number of guests which leads to better results. Apart from that while plotting the world map, we performed a left join between the Map\_data(preloaded data) and the hotel data set on the column ‘Country’ to remove the NA values.







**Data Transformation:**

For using the data in our models, we have first converted the variables to factors and stored them in a separate data frame. Furthermore, for Association rule mining, we converted this data frame to a transactional matrix.

Code Snippet of converting the variables into factors:

Text

Description automatically generated

# EXPLORATORY DATA ANALYSIS

Exploratory data analysis is the process of checking out the entire dataset and find relations between different variables in order to come to an actionable outcome. This is mainly done via analyzing the structure of the data and also creating visualizations which help in finding insights to the data which wouldn’t be apparent upon the inspection of numbers. For the purposes of the project, we carried out visualizations with the help of bar plots, boxplots, tables and scatter plots.

Bar plots:

Chart, bar chart

Description automatically generated

Here we can observe that customers who opted for bed and breakfast cancelled their booking the most (~7500 bookings were cancelled). Customers who opted for half mean tend to cancel their bookings 30% of the time.

Chart, bar chart

Description automatically generated

Here we can observe that customers who did a booking though an online travel agent cancelled their bookings the most (~6000 bookings). We can also see that customers who booked in groups cancelled their booking 40% of the time. We have good response from customers who made reservations directly though the hotel website and other direct channels.

Chart, bar chart

Description automatically generated

Here we can observe that we don’t have a lot of guests who rebook. (~11000) new guest have cancelled booking previously.

#We should start a loyalty programme as we can see that there are not a lot of guest who come back to stay at the hotel.

Chart, bar chart

Description automatically generated

Here we can observe that room A was reserved most of the time but 25% of those bookings were cancelled. Room D was the second most reserved room followed by room E. It is strange to see that people don’t reserve room B. Room L and P were the least opted room.

#Check the condition of room B, L and P and renovate the rooms to increase availability.

Chart, histogram

Description automatically generated

Here we can observe that most customers were assigned room. Almost 30% of those bookings were cancelled. Second highest assigned room is room D. room L, P and B were the least assigned room.

Timeline

Description automatically generated with low confidence

Here we can observe that most customers did not pay any deposit. Almost 30% of the times these bookings were cancelled. Interestingly we can see that guests who booked the non-refundable rooms cancelled there bookings more that 95% of the times.

Chart, bar chart

Description automatically generated

Here we can observe that most of our guests stay only for a short time (almost 80% of the guests). 1/3rd od these bookings were cancelled. We had a good number of transient-party as well. We have a low number of contract guests.

#Increase contract benefits to attract recurring income.

Box plots: Chart, box and whisker chart

Description automatically generated

Insight: when the difference between the booking date and arrival date is large, that is approximately more than 100 days, the customer tend to cancel it due change in plans or other reasons.

A picture containing scatter chart

Description automatically generated

Insight: Here the outliers suggests that more required car parking space affects the cancelation, as more space means less cancelations.

Chart

Description automatically generated with medium confidence

Insight: Even though the averages are same for both cancelled and not cancelled are same, the outliers suggest that the customers that have canceled previously tend to cancel the future bookings.

**Histograms:**

Chart, histogram

Description automatically generated

Here we can observe that most frequently occurring lead time was between 0-100days. We have almost 9000 guests who booked more than 100 days prior and less than 200 days. Around 5000 guests have reserved a room before 200 days but not more than 400 days.

Chart, histogram

Description automatically generated

Here we can observe that most guests did not have any special requests. Around 11000 guests made a special request and 5000 guests made more than 1 special requests.

**Tables of categorical variables:**

For EDA, we are using the table() command on some of the variables to show what data each of the column consists and what is the split amongst the different input values.

**table(df$Meal):**



From above result we can see that most preferred meal choice is the BB(Bed &Breakfast). The second most preferred meal type is HB(Half-Breakfast). There are some customers who have not specified their meal plans while the other tow options FB and SC have very low number of customer when compared to BB and HB.

**table(df$MarketSegment)**



From the above result we can see that a majority of the bookings have been done via an Online TA(travel agent).Bookings from Offline TA/TO is the next most common method of reservation. Coming in at 3rd and 4th we see that Direct bookings and Groups bookings are also a preferred choice among the customers. The 5th most common method for reservation is via Corporate bookings. And last but not the least, the hotel provides some complementary booking as well to the customers.

**table(df$ReservedRoomType)**



Here we can see that the most reserved room type is room A followed by rooms D,E,F,G,C. Rooms B,L and P have the least number of reservations out of all the available rooms.

**table(df$AssignedRoomType)**



From this result we can see and comparing it with the above result, we can see that people who reserved Room type A either got Room A or have been assigned another room. Same goes for the other available rooms in the resort as well. Although here we see that there are customers who have been assigned a Room I, which is not visible in the above result, indicating that it might not be preffered choice of the customers.

**table(df$DepositType)**



Here we observe that for a majority of the bookings, the customers have opted for No Deposit on the reservation. There are a few customers who have opted for a non-refundable reservation as well. Somehow the customers are least interested in refundable bookings, which is weird to begin with.

**table(df$IsRepeatedGuest)**



Here we see that the number of guests who are repeated customers for the hotel are much lower when compared to the first-time visitors to the hotel. Which means that customer retention rate of the hotel is quite low which they need to work on.

**table(data$CustomerType)**

Contract Group Transient Transient-Party

1776 284 30209 7791

Here we can observe that customer’s whose booking is not part of a group or contract and is not associated to other transient booking are highest, then comes the ones who has some association with transient-party. The customers who have some kind of contract associated are 1776 and ones with group are 284 respectively.

**table(data$Country)**

AGO ALB AND ARE ARG ARM AUS AUT AZE BDI BEL BGR BHR BHS BIH

24 3 5 11 57 2 87 210 3 1 448 5 1 1 1

BLR BRA BWA CAF CHE CHL CHN CIV CMR CN COL COM CPV CRI CUB

7 430 1 3 435 17 134 2 2 710 16 1 5 2 4

CYM CYP CZE DEU DJI DNK DOM DZA ECU EGY ESP EST FIN FJI FRA

1 8 27 1203 1 65 3 12 2 1 3957 33 151 1 1611

GBR GEO GGY GIB GRC HKG HRV HUN IDN IND IRL IRN ISL ISR ITA

6814 11 1 13 10 4 11 47 5 37 2166 5 6 28 459

JAM JEY JOR JPN KAZ KOR KWT LBN LKA LTU LUX LVA MAC MAR MDG

5 3 2 9 5 9 3 6 1 46 80 33 1 75 1

MDV MEX MKD MLT MOZ MUS MWI MYS NGA NLD NOR NPL NULL NZL OMN

6 6 1 2 6 1 2 10 10 514 123 1 464 14 11

PAK PER PHL PLW POL PRI PRT QAT ROU RUS SAU SEN SGP SMR SRB

4 1 16 1 333 9 17630 1 177 189 1 1 4 1 7

SUR SVK SVN SWE SYC SYR TGO THA TUN TUR TWN UGA UKR URY USA

4 12 11 304 1 1 1 6 1 23 12 1 23 8 479

UZB VEN VNM ZAF ZMB ZWE

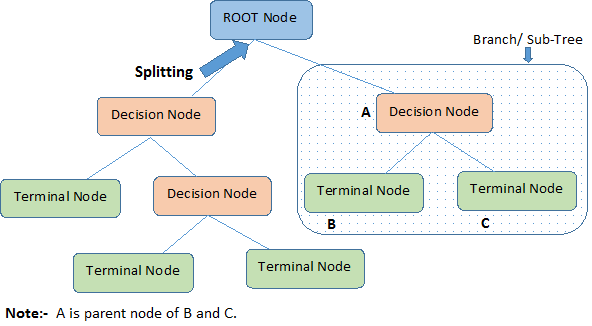
1 3 2 18 1 2

The greatest number of bookings are from Portugal and then Great Britain.

# CLASSIFICATION AND REGRESSION TREE

Decision trees are a form of supervised machine learning algorithm that can be used to solve a variety of classification problems and also regression problems. In the decision tree a question which is usually answered yes, or no is asked at each node and based on the answer, the algorithm moves down further to the next node where a new question is asked. The questions depend on the parameters or variables that are fed to the algorithm, and the parameter that provides the best split is selected for each node. The root node generally has the parameter that has the most correlation to the dependent or the target variable. The terminology used in a decision tree is given below:

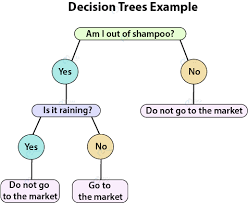
* Root node: this is the representation of an entire sample. At the top of the decision tree.
* Splitting: the process of separating a node into sub nodes. There can be two or more sub nodes.
* Terminal node: a node which cannot be split further.
* Pruning: the removal of a sub node of a node which has a decision.
* Branch: an entire section of a decision tree.
* Parent node: a node which divides further into sub nodes.
* Child node: the sub node which has come from a parent node.



The type of decision tree used in this project is a Classification and Regression Tree (CART). Whenever new data is fed to the decision tree, it starts at the root node and is moved forward based on the questions at each and every node until it reaches a terminal node. The dataset is divided into training set and a testing set. The training set contains a larger share of the dataset as it helps the model to learn more situations. The testing set is then used to test the model and the accuracy is measured using a confusion matrix, which makes a matrix to calculate the accuracy based on true negatives, true positives, false negatives and false positives.

The Advantages of using a classification and regression tree are:

* A decision tree is very easy to understand and interpret.
* It is very helpful in dealing with problems where the target variable, or the predicted value is a categorical variable.
* The data fed to a decision tree algorithm doesn’t necessarily have to be linear. It works well with non-linear data too.
* The tree is very simple as it just uses one variable for splitting at each node.



The R libraries used for the decision tree algorithm are rpart, e1071, rpart.plot and the caret library. The rpart library is used for the rpart() function which generates the decision tree based on the target variable and the features provided to it. The rpart.plot library is used for the rpart.plot() function which plots the decision tree for visualization. Finally the caret library is used for the confusionMatrix() function which gives the statistics and the confusion matrix generated from the algorithm and the test set provided to the predict() function. The predict function generates a matrix which consists of the predictions from the rows of input features in the test set.

The decision tree obtained by feeding the features LeadTime, PreviousCancellations, BookingChanges, RequiredCarParkingSpaces, TotalOfSpecialRequests, IsRepeatedGuest, DepositType, MarketSegment and AssignedRoomType, is shown in the figure given below:

A picture containing diagram

Description automatically generated

DECISION TREE

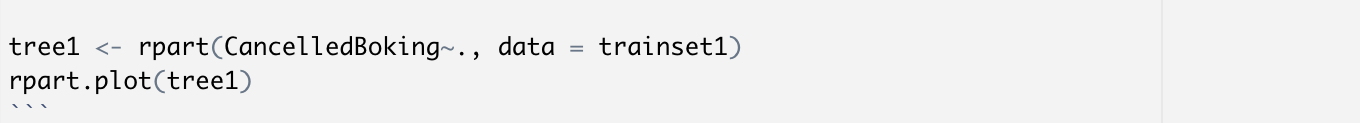
The decision tree was created using the following process:

1. Data transformation: The categorical variables were transformed to factors using the as.factor() function and then utilized in the model.
2. Train test split: The dataset was divided into a training dataset and a testing dataset. The former is used to train the model and for this purpose 60% of the dataset was dedicated to this. The remaining 40% of the dataset was dedicated for testing the model and its accuracy at predicting the outcome. The following snippet explains the train test split of the dataset.

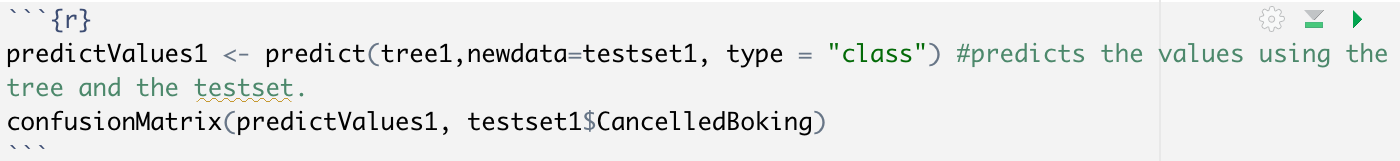
Graphical user interface

Description automatically generated with medium confidence

1. Tree modelling: The decision tree was generated using the rpart() function. This tree is an algorithm which is used for the prediction of the outcome in the testing data set based on the features fed to it. The rpart.plot() function is used to create a visualization of the decision tree. The following snippet explains the development of the decision tree algorithm:



1. Testing the prediction accuracy: The prediction of the model is generated using the predict() function form the caret package. This function generates a matrix which has the predicted outcome values for each row of testing dataset. The accuracy and statistics are obtained using the confusionMatrix() function. This function generates a confusion matrix and displays the statistics of the model based on the same confusion matrix. The following snippet shows the creation of the confusion matrix and the statistics based on it:



Based on the tree generated by the algorithm, the predict() function and the confusionMatrix() function give the following output:

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 10858 2323

1 717 2125

Accuracy : 0.8103

95% CI : (0.8041, 0.8163)

No Information Rate : 0.7224

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4678

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9381

Specificity : 0.4777

Pos Pred Value : 0.8238

Neg Pred Value : 0.7477

Prevalence : 0.7224

Detection Rate : 0.6777

Detection Prevalence : 0.8226

Balanced Accuracy : 0.7079

'Positive' Class : 0

According to the confusion matrix, where 0 stands for booking not cancelled and 1 stands for cancelled booking, the true positive value is 10858, that is according to the prediction total of 10858 bookings which weren’t cancelled were correctly predicted. The true negative value is 2125, that is the 2125 bookings which were actually cancelled were predicted correctly by the model.

Reference

Prediction 0 1

0 10858 2323

1 717 2125

The accuracy of the model, which is calculated by the formula given below, is 81.03%, which means that our model correctly predicts the booking cancellations 81 times out of 100.

The Sensitivity is the true positive values divided by the total positive values and it comes out as 0.93 which is really good. The Specificity is given by the formula:

The specificity comes out as 0.48.

The McNemar’s test gives the homogeneity of misclassification of the two target variable values, 1 and 0. The value is very low which is good for our model.

In conclusion, the decision tree algorithm with the abovementioned features is accurate 81% of the time.

# ASSOCIATION RULE MINING

Association Rule Mining is a set of If/Then statements through which we are able to identify patterns in the given data. Here we don’t consider the data as a table of rows and columns. Instead we view it as a set of transactions and what patterns do these transactions generate.

Parameters on which the relationships depend on:

* **Support:** It indicates how frequently an if/then relationships occurs in the dataset
* **Confidence:** It denotes that how much confident we are that a rule will hold true.
* **Lift:** Lift is a metric indicating how well a targeting model (association rule) predicts or classifies situations as having an improved response (in comparison to the population as a whole), as compared to a random choice targeting model.

To perform Association Rule mining, we will use the Apriori algorithm to first generate the set of rules

Apriori Algorithm:

This algorithm uses the prior knowledge of the properties of frequently occurring itemsets.This is an iterative algorithm/ level-wise search where n-frequent itemset will be used to find out n+1 itemsets.

Here’s how we proceeded with Association rule mining:

* First we created a dataset ‘hotel’ from the given .csv file.
* Then we are creating all the variables to factors and storing them in a new dataframe ‘hotel\_data’
* Further we convert the data frame into a transactional matrix.
* Now using the apriori function, we generate a set of rules while keeping LHS on default(which will consider all the variables in the matrix.) and RHS as IsCancelled=1, since we want to know the patterns/rules which ultimately lead to cancellations.
* Attached below are the code snippet and the set of rules that have been generated.

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

In the above screenshot, we can see that when the Market\_Segment= Groups, there tends to be a cancellation of the reservations. Furthermore,there is a rise in cancellations when the guests are from Portugal as well.(Country=PRT.)



To evaluate further, when we look at the above rule it states, that when Children=0, Country=PRT, Market\_segment=4, Booking\_changes=0, RequiredCarParkingSpaces=0, Special\_requests=0 we see that there is a higher chance of the reservation getting cancelled.

This means that apart from Country and Market\_segememt, other variables such as Children also play a significant part in deciding the possibility of the cancellation.

In the snippet below, we are performing the same steps as above, but instead of using all the variables, we are only using the key variables from the dataset.

Text

Description automatically generated

Graphical user interface, table

Description automatically generated

As we can see from the set of rules generated, whenever the Deposit\_type = Non refundable, there seems to be a higher chance of cancellations in the reservations. At same time we are getting higher values of confidence which means that although the occurrence of these combinations may be less but there is a higher chance of cancellations.

If we take a closer look at one of the rules for example rule 7 & 9. We can see that variables such as RequiredCarParkingSpaces, Special\_requests, booking\_changes have an impact on the probability of cancelation as well alongside the Deposit\_type.

# SVM: Support Vector Machine

Support Vector Machine finds a hyperplane that helps is classifying the data points. We can draw as many hyperplanes as possible to classify the data points. Our objective is to find a plane that has the maximum margin, i.e., the maximum distance between data points of both classes. We need to maximize the distance to provide some margin for the later data points that would be needed to classify.

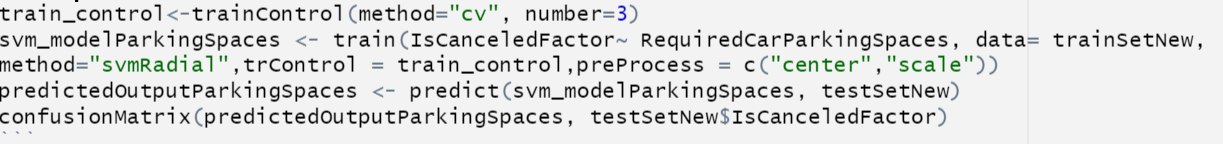
Chart, scatter chart

Description automatically generated

Dimensions of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane.

The points that are closer to the hyperplane are called as support vectors. The hyperplane is drawn based upon these support vectors. The support vector defines the margins and any change in these points can change the position of the hyperplane.

For our project, one of the models that we used to predict the cancellation of hotel was SVM. First, we began building the model using just one feature. But the accuracy was 72% or approximately around for most of them.



Output:

A screenshot of a computer

Description automatically generated with low confidence

Then we considered the combination of few features such as lead time, required car parking space, assigned room type etc. After considering these features we could increase the accuracy to 81.21%.

Here significant factors include lead time, booking changes, previous cancellations, required car parking spaces, total of special requests, is repeated guest, deposit type, assigned room type, and market segment.

Text

Description automatically generated

Output:

Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

Comparing the output of both we selected the significant factors and trained the model to get the accuracy of 81%.

# ACTIONABLE INSIGHTS

Based on the analysis of the data provided, we conducted exploratory data analysis using bar plots, box plots, tables and the descriptive statistics. We also compared all the machine learning models and came up with the following insights:

* *We should start a loyalty program as we can see that there are not a lot of guests who come back to stay at the hotel.*

This is based on the bar plot shown below. We can observe that we don’t have a lot of guests who rebook. (~11000) new guest have cancelled booking previously.

Chart, bar chart

Description automatically generated

* *Check the condition of room B, L and P and renovate the rooms to increase availability.*

We can observe that room A was reserved most of the time but 25% of those bookings were cancelled. Room D was the second most reserved room followed by room E. It is strange to see that people don’t reserve room B. Room L and P were the least opted room.

Chart, bar chart

Description automatically generated

* *Increase contract benefits to attract recurring income.*

We can observe that most of our guests stay only for a short time (almost 80% of the guests). 1/3rd od these bookings were cancelled. We had a good number of transient-party as well. We have a low number of contract guests.

Chart, bar chart

Description automatically generated

* *Use the decision tree model that was discussed earlier for the prediction of cancellations as it has the highest accuracy of 81.3% and has a much faster speed than the other models.*