Submitted By:

**GABRIELLA CAGGIANO | REETODEEP HAZRA | PRATEEK KUMAR KUMBAR | QIAOYI LIU | YASH SHAH**

**SCHOOL OF INFORMATION STUDIES, SYRACUSE UNIVERSITY**

**Data analysis for hotel reservations**

IST 687 Project Implementation Report

Lab Section M002 | group 5

Graphical user interface

Description automatically generated

Contents

[Description 3](#_Toc89972149)

[Project Scope and Objective 3](#_Toc89972150)

[Project Deliverables 3](#_Toc89972151)

[Data Acquisition 3](#_Toc89972152)

[Data Preprocessing 4](#_Toc89972153)

[Data Dictionary 5](#_Toc89972154)

[Data Visualization 8](#_Toc89972155)

[a) Histograms of categorical variables 8](#_Toc89972156)

[b) Reserved room types vs. assigned room types 9](#_Toc89972157)

[c) Independent variables vs. IsCanceled variable 9](#_Toc89972158)

[d) Boxplot of all numeric variables against IsCanceled 12](#_Toc89972159)

[a) Histograms of all numeric variables against IsCanceled 14](#_Toc89972160)

[e) Lead time and cancellations from different countries 16](#_Toc89972161)

[Modelling Techniques 17](#_Toc89972162)

[a) Decision Tree 17](#_Toc89972163)

[b) Association Rule Mining 19](#_Toc89972164)

[c) Support Vector Machines 19](#_Toc89972165)

[Key Findings 21](#_Toc89972166)

[Recommendations 21](#_Toc89972167)

[Future Scope 22](#_Toc89972168)

[Reference 22](#_Toc89972169)

# Description

The project evolves around analyzing the hotel data collected from several customers all over the world and using those data to provide business recommendations about why customers cancel their hotel reservations.

# Project Scope and Objective

The scope of this project is to analyze and draw insights from the dataset provided to us which contains data regarding hotel reservations based on various factors. The data contains data from several customers all over the world and contains 20 variables.

We will be concentrating on finding the factors that cause a customer to be unsatisfied and get actionable insights by applying statistical techniques.

The objective of this project is to understand some key factors for why people cancel hotel reservations. We also provided some insights on potential factors upon which customers are cancelling the reservations.

# Project Deliverables

* To prepare our data for further analysis by cleaning it and confirming that there are no missing or invalid fields in the dataset.
* Identify the key attributes that mostly affects customers to cancel their hotel reservation.
* We found out the factors using Support Vector Machine and Random Forest.
* Using Apriori algorithm to suggest rules that can be used to reduce the number of hotel cancellations.
* To formulate actionable insights after applying Support Vector Machine and Random Forest.
* Finally, to provide suggestions based on the data analysis to reduce the rate of cancellations.

# Data Acquisition

The dataset provided to us contains 40061 hotel reservation data with 20 fields such as IsCanceled, LeadTime, StaysInWeekendNights, StaysInWeekNights, Adults, Children, Babies etc.

Initially the data was extensively studied to determine some key variables. The dataset was then taken over to the preprocessing phase where the errors were removed and making it usable for further analysis.

# Data Preprocessing

Before preprocessing, the data set consisted of ​40061 rows and 20 column variables. All the data in the data set was the survey taken of customers making their hotel reservations.

Firstly, we checked for any null values in the dataset. There were 464 null values present in the Country column which were filtered out.

Secondly the columns were seen to determine any presence of error values and it was found that there were no error values present in the dataset.

Thirdly, we analyzed the number of people, i.e., Adults, Children and Babies. We figured out any observations that has 0 as total number of people. We found that there were 11 observations that contained 0 number of people. We removed those rows where total of adults, children and babies is 0 as the number of people staying at the hotel cannot be 0.

Fourthly, the total night stays in the dataset were analyzed. We found out that 376 rows have 0 for Stays in Weeknights and Stays in Weekend Nights. We filtered out those rows where total nights booked is 0 because this is a data entry error as no one can book a hotel for 0 nights.

Finally, we analyzed the meal preferences of the customers. There were 4 categories of meal booked; Undefined/SC which meant no meal package; BB which meant Bed & Breakfast; HB which meant Half board (breakfast and one other meal – usually dinner) and FB which meant Full board (breakfast, lunch, and dinner). So Undefined and SC both have the same meaning. So, we merged the two columns and named it SC.

After data munging the number of rows in the data set were ​39209 and had 22 columns.

**Code:**

#checking for null values in country column

table(data$Country) #464 rows have null in country column

data <- data %>% filter(Country!="NULL") #Removing the rows containing null values for country

#Analyzing number of people i.e., adults, children, and babies

data$People <- data$Adults + data$Children + data$Babies

data[data$People==0,] # 11 observations have 0 as the total no of people.

data <- data %>% filter(People!=0) #Removing those rows where total of adults, children and babies is 0 no. of people staying at the hotel cannot be 0.

#Analyzing total night stays

data$TotalNights <- data$StaysInWeekendNights + data$StaysInWeekNights

data[data$TotalNights==0,] #376 rows have 0s for both StaysInWeekNights and StaysInWeekendNights.

data <- data %>% filter(TotalNights!=0) #Removing those rows where total nights booked is 0 because this is data entry error as you cannot book a hotel for 0 nights.

#Analyzing the meal preferences of the customers.

table(data$Meal)

data$Meal[which(data$Meal == "Undefined")] <- "SC" #Since "Undefined" is SC, converting the Undefined to SC

view(data)

# Data Dictionary

|  |  |  |
| --- | --- | --- |
| Variables | Entity and Attributes | Description |
| Children | Integer | Total number of children |
| Country | Categorical | Country of origin represented in ISO 3155- 3:2013 version. |
| Meal | Categorical | Types of meal booked by customers. The packages are as follows:   Undefined/SC– no meal package   BB – Bed & Breakfast; HB – Half board (breakfast and one other meal)   FB– Full board (breakfast, lunch, and dinner) |
| IsCanceled | Categorical | Categorical Value which indicates whether the booking was canceled (1) or not canceled (0). |
| LeadTime | Integer | The number of days that passed between entering date and the booking date. |
| StaysInWeekNights | Integer | The number of weeknights the guest booked to stay in the hotel. |
| StaysInWeekendNights | Integer | The number of weekend nights (Saturday and Sunday) the guest booked to stay in the hotel. |
| Adults | Integer | Number of adults |
| Babies | Integer | Number of babies |
| MarketSegment | Categorical | Market segment description   * Online TA: Online Travel Agents * Online TO: Online Tour Operators * Direct Bookings * Corporate Bookings * Complimentary Bookings |
| IsRepeatedGuest | Categorical | Value indication whether the booking is from a previous guest (1) or not (0). |
| PreviousCancellations | Integer | Number of previous bookings that were cancelled by the customer prior to the current booking. |
| PreviousBookingsNotCancelled | Integer | Number of previous bookings that were not cancelled by the customer prior to the current booking. |
| ReservedRoomType | Categorical | Code of room type reserved like A,B,C,D etc. |
| AssignedRoomType | Categorical | Code of the room type assigned to the guest. |
| BookingCharged | Integer | Number of changes made by the guest after booking. |
| DepositType | Categorical | Indication on if the customer has made a deposit or not. There were three categories:   * No Deposit * Non Refund * Refundable |
| CustomerType | Categorical | Type of booking made by the guest. There were four categories.   * Group * Transient * Contract * Transient-party |
| RequiredCarParkingSpaces | Integer | Number of car parking spaces required by the guest. |
| TotalOfSpecialRequests | Integer | Number of special requests made by the customer. |

# Data Visualization

## Chart Description automatically generatedChart, bar chart Description automatically generatedHistograms of categorical variables

Bookings that had not been cancelled was twice more than cancelled.

Chart, bar chart

Description automatically generated

Most reservations have made no deposits.

Chart, bar chart

Description automatically generated

**Chart, bar chart

Description automatically generated**

Most customer preferred online travel agents to complete their booking.

Most reservations associated with the Bed & Breakfast meal plan.

Transient customer type outweighed the rest of the customer types.

## Chart, scatter chart Description automatically generatedReserved room types vs. assigned room types

The figure provides insights that the demand for room type A is the highest. But also, it interprets that the customers with room type A also tends to cancel the most.

## Chart, waterfall chart Description automatically generatedChart, waterfall chart Description automatically generatedIndependent variables vs. IsCanceled variable

From the above graph, we get an insight about the meal plan of the customers. Most customers reserve their hotel rooms with the Bed & Breakfast meal plan. There are almost no reservations with the Full Board meal plan.

From the statistics of this graph, 95% of customers tend to cancel the reservation despite the deposit type being non-refund.

**Chart, bar chart, waterfall chart

Description automatically generatedChart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart, waterfall chart

Description automatically generated**

This graph helps us to understand the market segment relation with cancellations. The customers who have booking with Online Travel Agents have the same ratio of bookings as well as cancellations. But the customers with Offline Travel Agents booking have very low number of cancellations. The same goes with the customers who are booking directly.

From this graph, 27.5% of customer who reserved room type A are more likely to cancel the reservation.

Repeated guests are less likely to cancel their reservations.

Most customers are transient and transient-party.

**Chart, waterfall chart

Description automatically generatedChart, bar chart

Description automatically generated**

Reservations with only adults are less likely to be canceled.

Same with reserved room type, reservations assigned to room type A are less likely to cancel.

## Chart, box and whisker chart Description automatically generatedBoxplot of all numeric variables against IsCanceled

The longer the lead time of reservation, the more likely the customer is to cancel. The rest of boxplot did not show a significant correlation between canceled and not canceled reservations. The rest of the box plots with other numerical variables are given below for reference

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated

Boxplot of StaysInWeekNights plotted with IsCanceled.

Boxplot of IsRepeated Guest plotted with IsCanceled.

Chart, box and whisker chart

Description automatically generated**Chart

Description automatically generated**

Boxplot of Previous BookingsNoCanceled plotted with IsCanceled.

Boxplot of StaysInWeekendNights Guest plotted with IsCanceled.

**Chart, box and whisker chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart

Description automatically generated**

Boxplot of TotalOfSpecialRequests plotted with IsCanceled.

Boxplot of RequiredCarParkingSpaces plotted with IsCanceled.

Boxplot of Previous Cancellations plotted with IsCanceled.

Boxplot of Babies plotted with IsCanceled.

Boxplot of Children plotted with IsCanceled.

Boxplot of Adults plotted with IsCanceled.

## Histograms of all numeric variables against IsCanceledChart Description automatically generatedChart, histogram Description automatically generatedChart, histogram Description automatically generatedChart, histogram Description automatically generated

Histogram between Adults against the frequency.

Histogram between StaysInWeekNights against the frequency.

Histogram between LeadTime against the frequency.

Histogram between StaysInWeekendNights against the frequency.

**A picture containing graphical user interface

Description automatically generatedChart

Description automatically generatedChart, histogram

Description automatically generated**

Histogram between BookingChanges against the frequency.

Histogram between PreviousBookingsNotCanceled against the frequency.

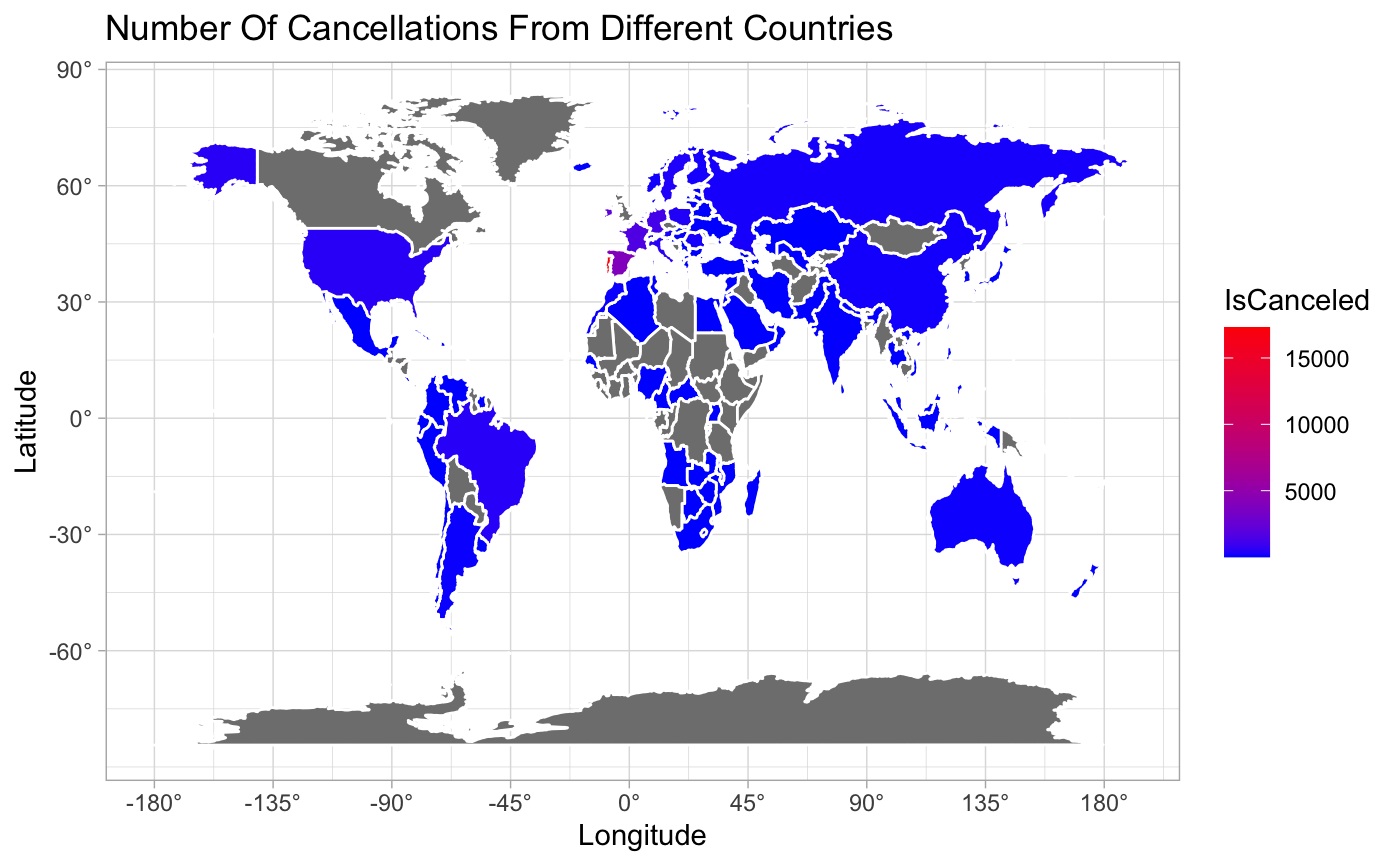
Histogram between PreviousCancellations against the frequency.

## Map Description automatically generated with medium confidenceLead time and cancellations from different countries

Countries with more lead time in reservations have darker color. From this map we can see that countries/region such as Iran, Columbia, Thailand etc. have more lead time and are more likely to cancel the hotel reservations.

**Map

Description automatically generated**



This world map is plotted to see the number of cancellations from different countries. We can see that most number of bookings and cancellations are made in the European countries. The country with most number of cancellations is Portugal which is shown in red color.

# Modelling Techniques

Various models have been used for accurate results of the information obtained from the data set. These models give an understandable representation of real-life information from the data sets. The following models have been implemented.

## Decision Tree

To begin with, we applied Decision Tree algorithm to our dataset. By applying this we could summarize and visualize the tree and see the most important variables by studying the root node and the parent nodes. ​

We found out several significant variables out of which we decided the top five significant variables based on further analysis of the model. As per our analysis, the following variables had a statistically significant relationship with the dependent variable in this project:

1. Lead Time
2. Required Car Parking Spaces
3. Market Segment
4. Country
5. Deposit Type
6. Total Nights

**Cart Tree:** The output of the cart tree using the rpart plot is shown below.

Graphical user interface

Description automatically generated

**Fig: Tree showing different nodes**

**Output:** The model output which gave us the most significant variables are listed here for reference.

Table

Description automatically generated

**Fig: Figure showing the most significant variables**

**Prediction and Accuracy:**

Table

Description automatically generated

The prediction of our decision tree model came as 82.26% and the recall came as 91% which is very good compared to other machine learning models. The recall calculates how many of the Actual Positives the model can capture through labeling it as Positive. So, 91% of the actual positives are correctly being calculated. The decision tree algorithm provides the best accuracy among other algorithms.

## Association Rule Mining

From the rpart plots, we found the top attributes affecting hotel reservations. Next, we used the Apriori Algorithm on the data set to get a more clinical knowledge about the top attributes.

We first visualized the rules which came by using the algorithm. The screenshot of a part of the output is shown below.

Table

Description automatically generated with medium confidence

**Fig: Figure showing the rules by Association Rule Mining**

Insight from the figure: We chose the rules which have lift values greater than 1.2 and have higher confidence and support to be the best rules.

## Support Vector Machines

We used the SVM modeling techniques to predict the reasons for canceling hotel reservations by using various significant variables from our model.

We divided the dataset into training and testing so that we can check and validate our results.



**Prediction and Accuracy:**

Text

Description automatically generated with low confidence

The overall accuracy of the SVM model came as 83%. of the confusion matrix is same as the confusion matrix. From the output we also checked the ‘No Information Rate’, which is 0.7 and it is the error rate when the input and output are independent.

Chart

Description automatically generated**Supporting Histogram:**

# Key Findings

From the visualizations and data modeling, we found out several key findings.

* From the Lead Time vs Is Canceled plots, we noticed that, more the lead time the more likely the customer would cancel their reservation.
* The guests who are coming from the first time are more likely to cancel than the repeated guests.
* By visualizing the customer type, we can interpret that the guests with Group bookings have the highest cancellation rate.
* The guests booking through non-refunded deposit type tends to cancel their reservation more often than the guests who booked through no deposit or refundable deposit types.
* Considering the meal plans chosen by the guests, those guests taking Bed and Breakfast meal plan has the maximum number of cancellations than the guests taking Full Board or Half Board meal plan.
* The dataset contained maximum data from the European countries and by producing a world map, we figured it out that the people from Portugal cancelled their reservations for maximum time.
* By looking at the reserved room types, we see that mostly all the customers wanted to stay in Room Type A. But consequently, the guests who made their booking in room type A also cancelled their reservations for most of the time.

# Recommendations

Based on the data visualization and modeling, we present the following suggestions to decrease cancellation:

* Although the cancellation rate in non-refundable deposit type is higher, there are more customers choosing no-deposit type. So, it is crucial to attract customers choosing no-deposit from cancelling their bookings. Linking to longer leading time leads to more cancellation, hotels should limit the lead time for no-deposit reservations.
* The Bed and Breakfast meal type, being the highest chosen meal type, has 26% chances of cancellation. Hotels should target this customer group with discounts to upgrade their meal plans from Bed to Breakfast to Full Board type or Half Board type.
* Since more customers are choosing online travel agents, hotels can try to reduce the brokerage charges to attract more customers.
* In this dataset, there are more customers from Portugal and European countries that are likely to cancel their reservations. Hotels should survey on these customers and try to increase their accommodation experience.

# Future Scope

* The decision tree algorithm gave us the best accuracy. We can further continue the analysis and improve the prediction of reservations cancellation by leveraging the reviews data with the help of text analysis.
* We can build models using state of the art deep learning techniques to further improve the accuracy of predictions.

# Reference

* Jeffrey S. Saltz and Jeffrey Morgan Stanton, “Data Science for Business With R”; ISBN 978-1544370453