**A Report on  
HOTEL INDUSTRY ANALYSIS**

**By  
Ankita Vartak (SUID: 358554918)**

**Nivesh Vaze (SUID: 411946624)**

**Mikhail Pinto (SUID:** **321933957)**

**Dhrumil Parekh (SUID:** **461973891)**

Table of Contents

1. Project Description1
2. Project Technical Details1
3. Project Goal2
4. Objective2
5. Packages Required3
6. Data Importing and Cleaning4
7. Exploratory Analysis9
   1. Basic Plots9
   2. Box Plots11
   3. Bar Plots19
8. Geographic Visualization25
9. Models34
10. Recommendations46
11. **PROJECT DESCRIPTION:**

* Hotel Industry has always been volatile when it comes to reservation and cancelation of rooms. Bookings for hotel rooms are done days and months before the arrival date which can sometimes lead to cancellation of bookings.
* Major factors for cancellations are assigned room type, seasonality, days of week etc. Thus, for a hotel industry it is very important to understand the historical data and analyze the trends which affects the hotel booking. These factors can be used for reporting the trends and predict the future bookings.
* To implement these, we will perform some essential analyses on dataset and generate the insights.

1. **PROJECT TECHNICAL DETAILS:**

* There are 20 features and 40060 observations available in the dataset.
* The dataset is not overwritten, and columns are not renamed for ease of understanding and coding.
* Checked the Missing and Null values in the dataset.
* Identified the categorical and numerical data separately to ease conversion of categorical data to numerical data.
* Encoded the vectors as factors for columns having less distinct values.
* Displaying unique values of all the categorical data.

1. **PROJECT GOAL:**

The overall goal of the project is to provide actionable insight, based on the data available.

1. **OBJECTIVE:**
2. To perform analysis such as: cancellation analysis, customer type, market segment analysis, among others.
3. Develop Machine Learning algorithms to predict cancellation by customer segmentation and days of week.
4. To analyze why people cancel hotel reservations and predict who will be cancelling
5. To analyze:
6. The number of cancellations:

* Number of bookings on weekday vs weekends
* Most preferred meal types
* Country wise bookings
* New customers acquired
* Type of rooms preferred by customers
* Booking types
* Assigned Rooms
* The number of guests in each booking

1. We will analyze patterns associated with each segment such as:

•  Day of week

•  Type of customers

•  Type of rooms

•  Market Segment

1. We will predict the future cancellation based on various machine learning algorithms such as Apriory algorithm, linear modelling and support vector machine.
2. Using these results, we can make key business decisions regarding the customer experience they desire to deliver.
3. **PACKAGES REQUIRED:**

* tidyverse- The tidyverse is a collection of R packages
* ggplot2- It is used for data visualization
* RCurl- It is an R-interface to the libcurl library and provides HTTP facilities
* maps- Provides different map outlines and points
* ggmap- Provides with functions to visualize spatial data and models
* mapproj- Converts latitude/longitude into projected coordinates
* countrycode- Converts to and from several different country coding schemes.
* Rworldmap- For mapping global data
* imputeTS- Specializes in time series imputation
* caret- Used for building machine learning models
* kernlab- Used for kernel-based machine learning methods in R
* rpart- Rpart is a powerful machine learning library in R that is used for building classification and regression trees
* randomForest- Random Forest in R used for classification and regression
* arules- The arules package for R provides the infrastructure for representing, manipulating and analyzing transaction data and patterns

**Questions:** Based on the given data, we can analyze and find answers to the following questions:

1) What are the trends for hotel bookings?

2) How long do people stay?

3) How often do people cancel their booking?

4) Do people cancel more on weekdays or weekends?

5) What are the room and meal types preferred by customers?

6) Overview of cancellation rate w.r.t countries

7) What are analysis on customer type of booking?

1. **Data Importing and Cleaning:**

library(tidyverse)

hotel\_data <- read\_csv("https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv")

view(hotel\_data)

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**Observations**:

* The hotel\_data data set has 20 columns and 40,060 records.
* From all the records, 23399 records having reserved room type as “A”. Thus, we can say that room type “A” is preferred by most of the guests.
* Guests stay longer during weekdays than weekend.
* Most of the guests prefer to not to pay any deposit.
* Out of all the bookings, 30209 records have customer type as transient. Thus, we can say most of the guests are neither associated with any group, nor associated with any transient booking.
* Out of total cancellations (4075), most of them are done by non-repeated guests (3844).
* Country value is NULL in 464 records.

Summary and Structure of Data Set –

str(hotel\_data)

summary(hotel\_data)

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

1. **Exploratory Analysis:**

For exploration of the data set, we have created few basic plots of certain key columns which gives us an overview of the data and columns in a factored way. Below are the plots along with a few lines for explanation of each.

* 1. **Basic Plots**

7.1.1 Market Segment

Chart, bar chart

Description automatically generated

From the above plot, it is clearly visible that the market segment is highly dominated by Online Travel Agents, followed by offline travel agents and direct booking. Corporate bookings are least frequent.

* + 1. Customer Type

Chart, waterfall chart

Description automatically generated

The maximum customer/bookings are Transient in nature. Out of total 40k records, approx 30k(67%) records have Customer Type as Transient.

* + 1. Deposit Type

Chart, waterfall chart

Description automatically generated

It is clearly visible from the plot that almost all the bookings are done without paying any deposit. To be specific, approx 85% of bookings were made without paying any deposit.

**7.2 Box Plots**

Further, we have plotted boxplots to look for outliers in the data set.

* + 1. Boxplot of count of childrens per booking

A picture containing table

Description automatically generated

The above plot gives us an overview of the number of childrens per bookings. As we can see, there is a booking with 10 or more childrens. All other bookings have upto 4 childrens.

* + 1. Boxplot of count of Adults per booking

Table

Description automatically generated

The above boxplot is self explanatory. The distribution of count of adults per booking is varied from 0-50. Majorly the number of adults per booking is below 5.

* + 1. Boxplot of lead time per booking

Chart, box and whisker chart

Description automatically generated

From the above plot, it is visible that the normal leadtime of all the bookings is mostly upto 200 days with few bookings around 600 days of leadtime.

* + 1. Boxplot for Stay on weeknights

**Table

Description automatically generated**

* + 1. Boxplot for Stay on weekends

Chart

Description automatically generated

The plot of stay on weekday and stay on weekend suggests that people mostly stay in this hotel on weekday.

Table of categorical responses for analysis: -

1. Table of reserved room type

table(hotel\_data$ReservedRoomType)

table(hotel\_data$DepositType)

table(hotel\_data$CustomerType)

Graphical user interface, text, application

Description automatically generated

7.2.6 Boxplot for Cancellation data w.r.t Repeated Guest

boxplot (IsCanceled ~ IsRepeatedGuest, data = hotel\_data, xlab = 'Repeated Guest', ylab='Cancellation Data', col='red')

Chart, bar chart

Description automatically generated

The above plot suggests that the repeated guest do not tend to cancel the booking. Thus, hotel management must try to provide incentive to non-repeated guests so that they may continue

7.2.7 Boxplot for Cancellation Data w.r.t Market Segment

boxplot (IsCanceled ~ MarketSegment, data = hotel\_data, xlab = 'Market Segment’, lab='Cancellation Data', col='green')

Chart, bar chart

Description automatically generated

The above plot suggests that the cancellation of the booking is done mostly by the Groups and Online Travel Agents.

7.2.8 Boxplot for Cancellation Data w.r.t. Customer Type

boxplot (IsCanceled~ CustomerType, data=hotel\_data, ylab='Cancellation Data', col='blue')

Chart, bar chart

Description automatically generated

The above plot suggests that the cancellation is done majorly by the Transient customer type.

7.2.9 Weekend nights w.r.t customer type and room type

Chart

Description automatically generated

Above boxplot for bookings in weekend nights according to customer type and reserved room type. Here we can infer that contract-based customers prefer only A, B, C, D types of rooms. Where C, D rooms are booked maximum for weekend nights. Group based customers prefer A, B, C, D, E, F, G with G the least on weekend nights. Transient customers prefer A, C, D, E, F, G, H, L, P WITH B, L, P the least. Transient-Party customers prefer A, B, C, D, E, F, G. There are some outliers in all customer types for weekend nights

7.2.10 Weeknights w.r.t customer type and room type

Chart

Description automatically generated

Above boxplot for bookings in weeknights in each market Segment with assigned room type. Here we can infer that customer from Direct market segment have highest room bookings for weekend nights. Here different colors are represented according to Room types. The median is 5 in complementary market for L type room bookings in weekend nights bookings

Chart

Description automatically generated

Above shows distribution of babies where we can infer that majority of guests don’t have babies.

* 1. **Bar Plots**

Chart, histogram

Description automatically generated

The above plot compares the room type with respect to the cancellation data. It is clearly visible that room type A is the clear preferred irrespective of cancellation data.

Chart, bar chart

Description automatically generated

The above plot gives the comparison of market type with respect to cancellation data. Online TA market segment contributes most to the cancellation.

The below three grouped bar chart shows the comparison of deposit type, assigned room type, and market segment with respect to cancellation data.

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

This graph gives us the average value of IsCanceled vs all unique values of Lead Time. From this scatter plot it can be seen that for some of the values in Lead time the average value of IsCanceled is above 0.5, this indicates that particular unique value in LeadTime has a lot of cancellations. In future it can be marked as an indicator.

Chart, scatter chart

Description automatically generated

Similarly in this graph we have the same scenario but in this graph we evidently see that the scale is upto 0.6 which means that we cannot deduce according to this as there are extremely low amount of unique values in StaysInWeekendNights

Chart, scatter chart

Description automatically generated

Similarly in this graph we have the same scenario but in this graph we evidently see that the scale is upto 1 which means that we can deduce (not certainly) according to the unique values in StaysInWeekNights for which bookings are cancelled.

1. **Geographical Visualization**

**Map 1**

hotel\_data\_Canceled\_1 <- subset(hotel\_data, IsCanceled == 1)

hotel\_data\_Canceled <- subset(hotel\_data\_Canceled\_1, Country !='NULL')

dev.new()

map1\_data <- aggregate(x= hotel\_data\_Canceled$IsCanceled, by = list(hotel\_data\_Canceled$Country), FUN = sum)

sPDF <- joinCountryData2Map( map1\_data

                             ,joinCode = "ISO3"

                             ,nameJoinColumn = "Group.1",

                             nameCountryColumn = "Country")

mapCountryData(sPDF, nameColumnToPlot="x", colourPalette = "heat" ,addLegend = TRUE, aspect = 1, missingCountryCol = NA, add = FALSE)

Map

Description automatically generated

Using the recommended functions joinCountryData2Map() and mapCountryData() we can plot various graphs according to the Country.

In the map above, firstly, the data is filtered using conditions IsCanceled ==1 which indicates cancelled booking and cleaning the data by filtering out NULL present in the Country column of the database.

Secondly, using the aggregate feature which can be used to perform mathematical expressions and then grouped by a particular column content. In this case it is taking the sum of the cancelled bookings grouping them by country.

Finally, the data of world map and its polygon co-ordinates are merged with our data set using the abbreviations for each country. There are some which we are unable to plot is because of there is no such country abbreviation which matched to the world map data. This data is finally mapped, and we get the above results giving us the countries with least cancellations to the most by heat color palette.

Table

Description automatically generated with low confidence

From the above map we can see which countries customers are likely to cancel their bookings.

**Map 2**

dev.new()

map2\_data <- aggregate(x= hotel\_data\_Canceled$StaysInWeekendNights, by = list(hotel\_data\_Canceled$Country), FUN = sum)

sPDF1 <- joinCountryData2Map( map2\_data

                             ,joinCode = "ISO3"

                             ,nameJoinColumn = "Group.1",

                             nameCountryColumn = "Country")

mapCountryData(sPDF1, nameColumnToPlot="x", colourPalette = "heat" ,addLegend = TRUE, aspect = 1, missingCountryCol = NA, add = FALSE)

Map

Description automatically generated

On Similar lines as the above-mentioned map, we use aggregate function to take a sum of nights stayed on weekends grouped by the country.

Table

Description automatically generated with medium confidence

**Map 3**

dev.new()  
map3\_data <- aggregate(x= hotel\_data\_Canceled$StaysInWeekNights, by = list(hotel\_data\_Canceled$Country), FUN = sum)

sPDF3 <- joinCountryData2Map( map3\_data

                              ,joinCode = "ISO3"

                              ,nameJoinColumn = "Group.1",

                              nameCountryColumn = "Country")

mapCountryData(sPDF3, nameColumnToPlot="x", colourPalette = "heat" ,addLegend = TRUE, aspect = 1, missingCountryCol = NA, add = FALSE)

Map

Description automatically generated

This map gives the sum of nights stayed on weekdays grouped by the country.

Table

Description automatically generated with low confidence

**Map 4**

dev.new()

map4\_data <- aggregate(x= hotel\_data\_Canceled$BookingChanges, by = list(hotel\_data\_Canceled$Country), FUN = sum)

sPDF4 <- joinCountryData2Map( map4\_data

                              ,joinCode = "ISO3"

                              ,nameJoinColumn = "Group.1",

                              nameCountryColumn = "Country")

mapCountryData(sPDF4, nameColumnToPlot="x", colourPalette = "heat" ,addLegend = TRUE, aspect = 1, missingCountryCol = NA, add = FALSE)

Map

Description automatically generated

The above map sums up the booking changes for cancelled booking by country names.

Table

Description automatically generated with low confidence

1. **Models**

**Linear Regression**

linOut\_1 <- lm(IsCanceled ~ LeadTime , data = hotel\_data)

summary(linOut\_1)

Text

Description automatically generated

In this Linear regression model the Adjusted R-squared value is extremely low which translates to low accuracy which is around 5.2% in this case. Even though, the Intercept and the variable itself is significant the accuracy is still low.

This issue is caused mainly because of the huge amount of data set. As we are taking only 1 column to predict another, the accuracy is low.

linOut\_2 <- lm(IsCanceled ~ StaysInWeekendNights, data = hotel\_data)

summary(linOut\_2)

Text

Description automatically generated

Like the above cases both intercept and variable are significant, but the accuracy is just 0.6%.

linOut\_3 <- lm(IsCanceled ~ StaysInWeekNights, data = hotel\_data)

summary(linOut\_3)

Text

Description automatically generated

Like the above cases both intercept and variable are significant, but the accuracy is just 0.6%.

linOut\_4 <- lm(IsCanceled ~ Adults, data = hotel\_data)

summary(linOut\_4)

Text

Description automatically generated

Like the above cases both intercept and variable are significant, but the accuracy is just 0.6%.

By this we can say that the Linear Regression model in this scenario as the accuracy is turning out to be very low. We can move to multiple regression where we use more than one column to predict.

**Multiple Regression**

linOut <- lm(IsCanceled ~ ., data = hotel\_data)

summary(linOut)

Graphical user interface, application

Description automatically generated

Graphical user interface, table

Description automatically generated

We can observe when we select all the columns for prediction the Adjusted R-Squared value is 0.3831 which generally translates to 38.31% accuracy which is not very good, we can try to maximize the accuracy by keeping the factors which are significant. In this model we can see not many countries are significant in gauging the if the booking is going to be cancelled. We can eliminate counties from out next model. Similarly, we can also eliminate Meals, MarketSegment, Babies and CustomerType as they are not very significant in this model.

linOut1 <- lm(IsCanceled ~ LeadTime + StaysInWeekendNights + StaysInWeekNights + Adults + Children + IsRepeatedGuest + PreviousCancellations + PreviousBookingsNotCanceled + ReservedRoomType + AssignedRoomType + BookingChanges + DepositType + RequiredCarParkingSpaces + TotalOfSpecialRequests , data = hotel\_data)

summary(linOut1)

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Even though we eliminated some of the insignificant variables we still see that the accuracy keeps on dropping this is due to the linear modelling. In linear prediction models the number of inputs (predictors) to the model also play a big role in the accuracy/ adjusted R-squared value. As we reduce the number of inputs even though all the variables are significant, we can see that the accuracy goes down considerably. R-squared value 0.2266 which translates to 22.66% accuracy of this model. To have better insights using linear modelling we can subset data according to categories and try to maximize accuracy of the model which in turn will translate into better prediction model.

**SVM Model**

set.seed(11)

trainList1 <- createDataPartition(y=hotel\_data\_fac$IsCanceled,p=.50,list=FALSE)

trainSet1 <- hotel\_data\_fac[trainList1,]

testSet1 <- hotel\_data\_fac[-trainList1,]

svmModel\_1 <- ksvm(IsCanceled ~ ., data=trainSet1, C=5, cross = 3, prob.model = TRUE)

svmModel\_1

svmPred\_1 <- predict(svmModel\_1, newdata=testSet1)

table(svmPred\_1, testSet1$IsCanceled)

sum(diag(table(svmPred\_1, testSet1$IsCanceled)))/sum(table(svmPred\_1, testSet1$IsCanceled))

confusionMatrix(svmPred\_1, testSet1$IsCanceled)

In the above mentioned SVM model all the variables were factored. 30% of the data in the data set was used for training and rest for testing. Even by changing the numbers for training and testing the accuracy of the model varied 0.5%. The accuracy of the model is 87.18% which is quite higher than the ones achievable in Linear or Multiple regression models. We can safely say that we can use this model to predict if a booking is going to be cancelled or not. In this we can feed all the values and predict using predict() function, if the person might or might not cancel the booking.

We can also use Association rule mining to check the exact rules where and when the most cancellations occur.

Graphical user interface, text, application

Description automatically generated

**Unsupervised Machine Learning - Association rule mining**

Table

Description automatically generated with medium confidence

Apriori Algorithm

#if meal = FB, marketsegment = groups, depositType= non refund there is high chance of cancelations

#if meal=FB, assignedroomtype=A,depositType=non refund then support is 0.007 and confidence = 0.97 which tells us that there is high chance of cancellations

1. **Recommendation**

1) Provide incentives/benefits to non-repeated guests so that they do not cancel the booking. We can start with giving discounts on 2nd bookings for new/non-repeated customers. The drive of getting a discount for their next booking might stop them from cancelling the current booking.

2) Extra benefits to customers who pay early deposits. For ex. Free Breakfast meals. Deposit is a way of assurance and we must try to encourage each guest to pay a deposit to ensure the occupancy.

3) Providing group discounts to customers could be usefull is decreasing the probability of group cancellation.

4) Free room upgrade for repeated guests as per the availability. Also we can start implementing loyalty programs and rewards for repeated guests.

5) Room type A and D are cancelled the most. Considering the cancellation factor, we can strat overbooking these specific room types to ensure maximum occupancy and better revenue generation.

6) Throughout the historical data, we have seen as more the number of special requests lesser are the chances of cancellation. We can start understanding the small but special requests of each booking and make sure of providing the customized service to each customer. Better the services offered, better chances of converting the guest into repeated guest.