IST 687- Applied Data Science
Final Project Report

Hotel Booking Cancellation Prediction

SYRACUSE UNIVERSITY · FOUNDED 1870 ·

SUOS
CULTORES
SCIENTIA
CORONAT

By-
Ameya Mangaonkar - 601634662
Sayali Sant - 747787412
Prajakta Mane - 867969039

TABLE OF CONTENTS

1. Abstract

The following report is presented for the Hotel-Booking cancellation prediction project. It is critical to identify the factors that are responsible for the cancellation of hotel reservations. The given dataset for this project consists different variables against cancellations. The main aim is to identify the factors that impact the cancellations and predict who will cancel based on the different factors. We have decided to use four models for the 'IsCanceled' variable against all the other variables given to establish a relationship. The four models are as follows:

1. Support Vector Machine (SVM)
2. Linear Regression & Multiple Regression
3. Apriori Algorithm
4. R-Part Decision Tree

2. Exploratory Analysis

Dataset and attribute explanation

The original dataset provided to us contains real-life hotel information and the parameters associated with it. The most important columns provided to us contain details such as:

1. TotalNumberofNightsStayed: We have combined two columns viz. StaysInWeekendNights and StayInWeekNights to form one single column as stated above
2. MarketSegment: Defines the way customers have booked the accommodations. There are a total of four way through which the bookings have been made:

a. Direct: Directly by contacting the hotel
b. Corporate: through office bookings for business meetings
c. Online/Offline TA/TO: Bookings made through Online/ Offline Travel agents/ Tour Operators

3. DepositType: States whether a deposit was made to secure the booking. It contains three types:

a. Non-refund
b. Refundable
c. No-deposit

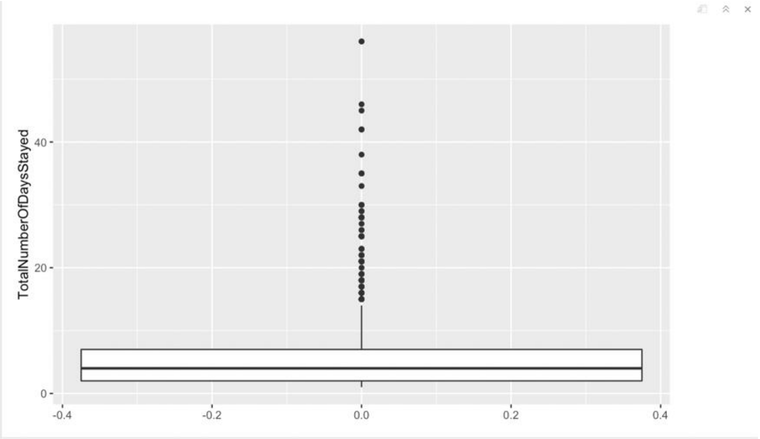
4. RequiredCarParkingSpaces: Provides us with the number of car parking spaces needed by a particular guest/customer
5. PreviousBookingsNotCanceled: States the exact count of previous bookings not canceled by the customer
6. CustomerType: Contains four types:

a. Transient: Staying only for a short time
b. Contract:
c. Group
d. Transient-party

1. Data Cleaning

Before performing any kind of analysis on the data, we initially cleaned the data, i.e., searched for NULL values or values that were insignificant and performed the following steps for cleaning the data. The steps were as follows:

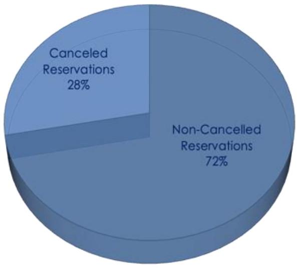
a. We removed the rows where the total number of nights stayed by the customer was 0
b. We removed the rows where the country was NULL
c. We removed the rows where adults were zero and children/babies were more than 1
d. The following boxplot explains that most of the customers do not stay for more than 15 days, and it makes no logical sense for customers to stay in a hotel for more than 15 days. Hence, we removed the rows where the total number of staying days were more than 15.



2. Initial Insight

According to our initial findings, we found that 28% of the people cancel their hotel bookings. This is just an initial analysis, not based on any factor/attribute. This is a significant figure for a hotel that must operate effectively in order to maintain its development, and the issue must be addressed.

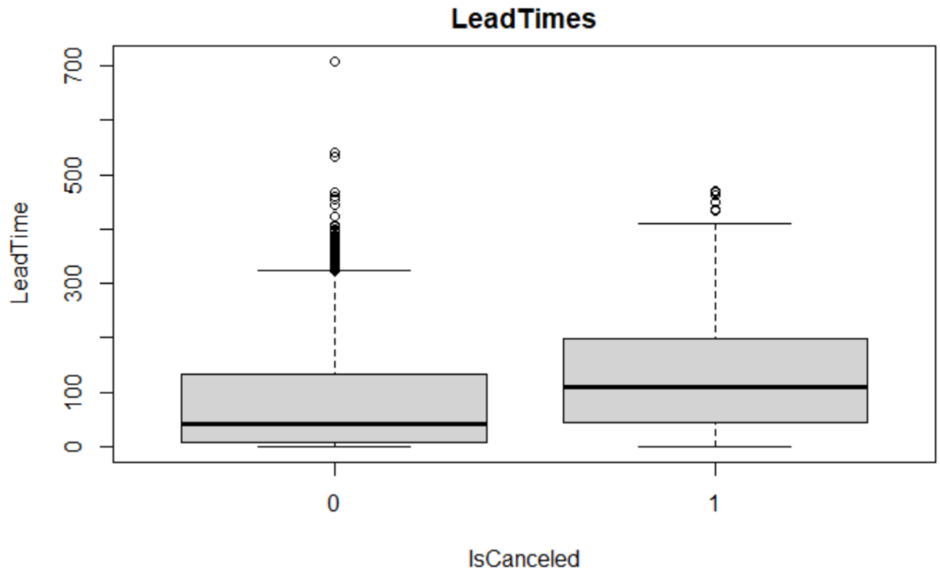
Code: AverageCancellationRate<-mean(df2$IsCanceled)



3. LeadTime Analysis

How many days ago was the booking made?

We found out from the boxplot that most bookings were made between 0-120 days which is 4 months. According to the following boxplot, we inferred that the greatest number of cancellations were when the **Leadtime** was between 50-120 days. We also found the median was about 40 days when there were no cancellations, and the median was about 100 days when bookings were cancelled. We also assumed that if the **Leadtime** was more, that means people would cancel their booking. But that was not true because we found out that people did not cancel their bookings when the leadtime was greater than 300 days.

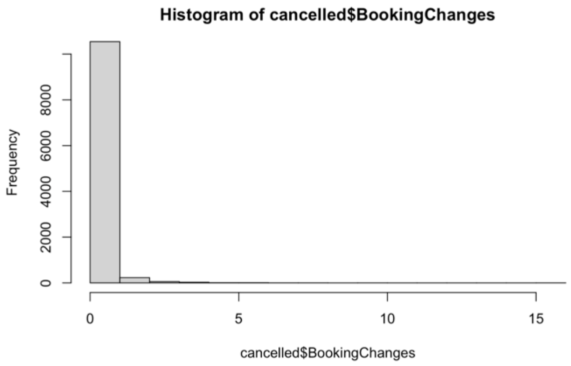


4. Booking Changes Made

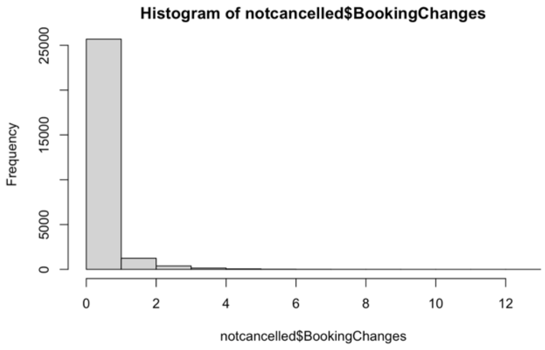
Here we attempt to find out if there was a change in the booking made by the customer until the moment he checks in at the hotel.

From this histogram, we inferred that more than 25000 customers have not changed the status of their booking and ended up checking in at the hotel while around 2000 customers have changed their booking status at least once before checking in.

On the other hand, around 10000 people have not made any changes to bookings and ended up cancelling their bookings. While around 500 people have made changes to their bookings before cancelling. We can ask these customers for the reason behind cancellation as that might help us to give insights into the requests of the customers and we can try to fulfil the demands of our future customers.



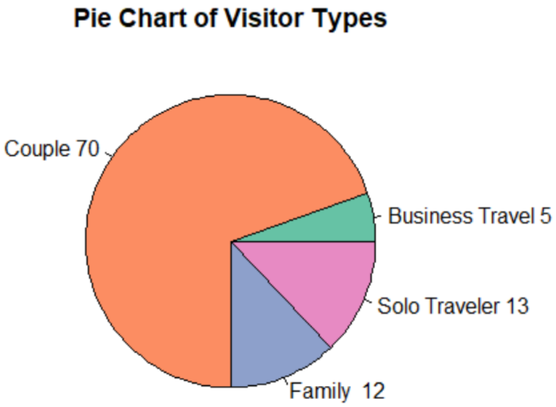
BOOKING CANCELLED



BOOKING NOT CANCELLED

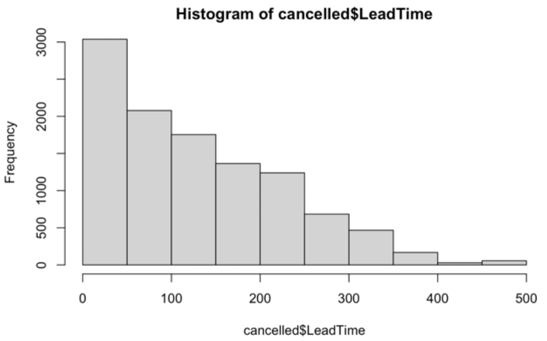
5. Pie Chart by Visitor type

To find out the percentage of the category of customers, we have created a pie chart which depicts the following findings:

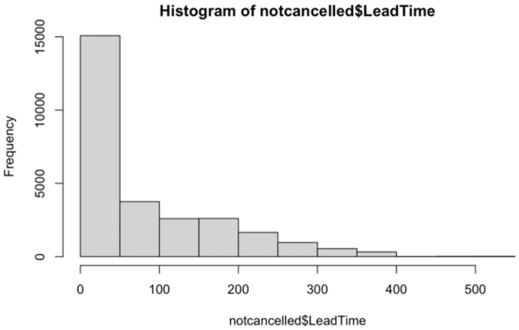


Here we can observe that couples visit the hotel and occupy up to 70% of the visitor type category. That is followed by Family which constitutes 12% of the visitor type category. The business travel customers and solo travelers constitute 5% and 12% respectively.

6. Lead Time Histogram:



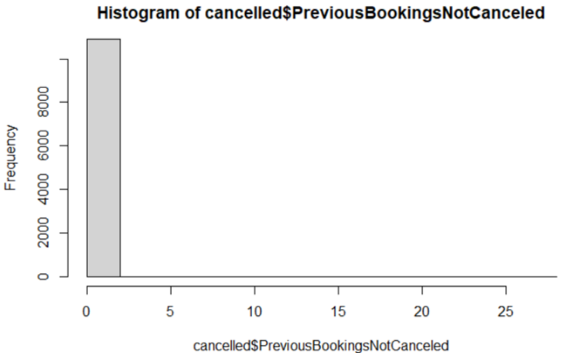
BOOKING CANCELLED



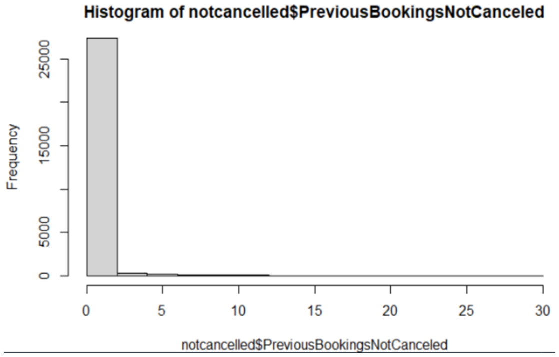
BOOKING NOT CANCELLED

From the above histogram, it is observed that the maximum cancellations occur when the lead time is between 0-100 and the frequency is around 3000 while maximum bookings are not canceled when the lead time is around 0-100 and the frequency is around 15000 which is 5 times the number of bookings canceled. So, based on this we can say that most people will not cancel their booking when the lead time is between 0 and 100.

7. Previous Bookings not Canceled:



BOOKING CANCELLED

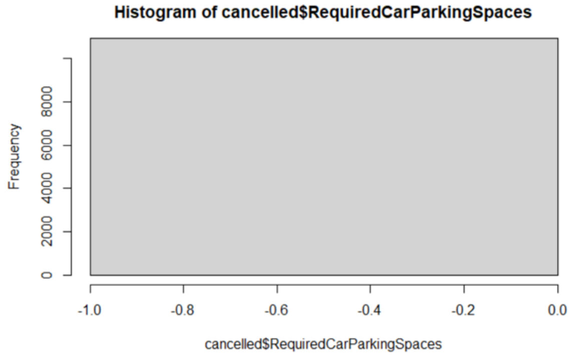


BOOKING NOT CANCELLED

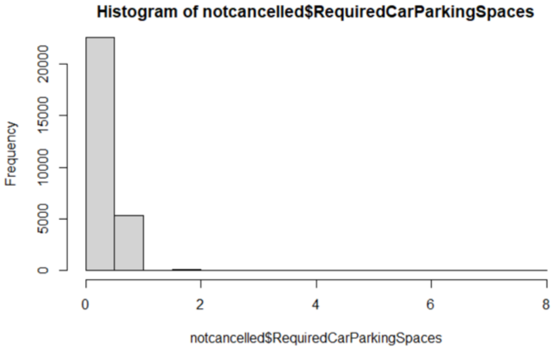
From the histogram, we can infer that more than 25000 people do not cancel their booking when their previous bookings at the hotel are not canceled. This means that these customers are persistent and really likely to stay at the hotel and won't cancel their future booking at the hotel.

On the other hand, more than 10000 customers have canceled their booking even though they hadn't canceled their previous bookings at the hotel. This means that those customers might have had some issues with the past experience in the hotel which needs to be known and analyzed

8. Required Car Parking Spaces:



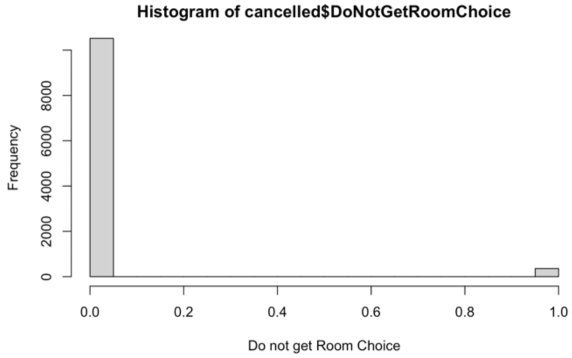
BOOKING CANCELLED



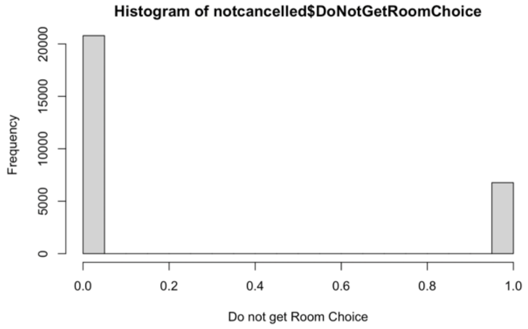
BOOKING NOT CANCELLED

From the above histogram, we can infer that 100% of customers cancel their booking when they do not find a parking spot and more than 20,000 customers stick to their booking when they at least find one parking spot and more than 5000 customers do not cancel their booking when they find more than 2 parking spaces. Through this, we can learn that it's important to provide car parking for the customers.

9. Do not Get Room Choice



BOOKING CANCELLED



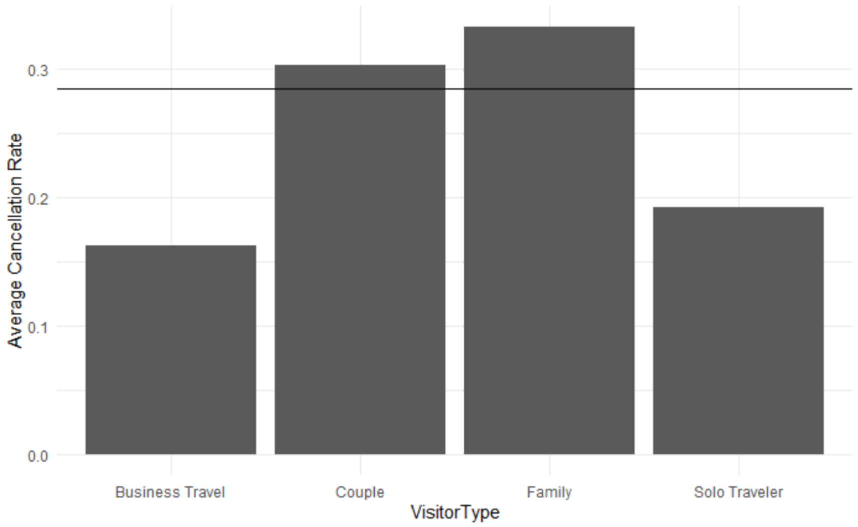
BOOKING NOT CANCELLED

From the above histograms we can observe that in total around 10000 people do not get the rooms of their choice but out of those 10000 people only 500 end up cancelling the bookings. Providing the customers with rooms of their choice which they had mentioned while booking will avoid these cancellations.

10. Bar plot for Visitor types

The average rate of cancellation is 28% and from the bar graph we can see that Couple and Families cross the average rate of cancellations.

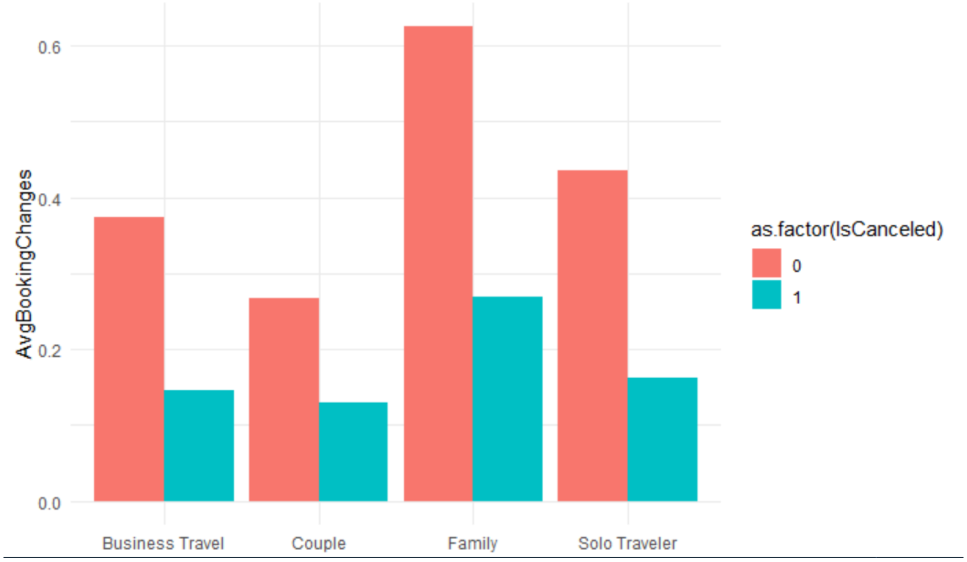
From the following bar plot, we can predict that families depict the highest rate of cancellation, followed by couples and solo travelers. Business travelers have depicted the lowest rate of cancellation. So, there is a probability for families to cancel their bookings in the future. We need to understand the reason behind the booking can cancelation and the peak period when the families cancel their bookings and accordingly provide offers to them so that we can restore their bookings



11. Booking changes Analysis

From this bar plot, we can infer that families have the highest rate of making changes to their bookings as compared to couples that have shown the least rate of making changes to their bookings.

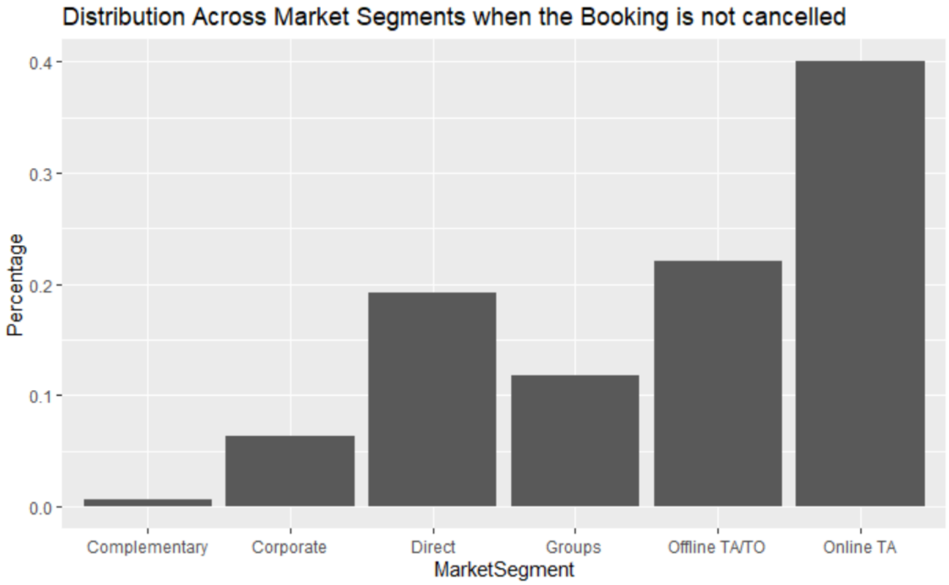
Another insight that we gained was that families had the highest chance of canceling their bookings when there were changes made to their bookings. Alternatively, couples had the lowest chance of canceling their bookings when there were changes made to their bookings.



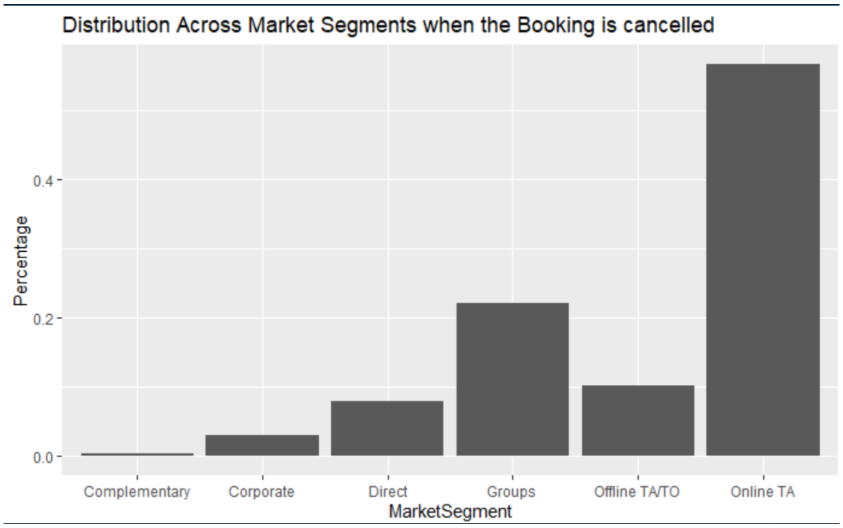
12. Market Segment Analysis

Most of the booking for the Hotel are made through Online Travel Agent mode which is followed by Offline Travel Agent mode.

From the first graph we observe that almost 40% of the hotel bookings that have been made through Online Travel Agents tend to not get cancelled.



From the following graph we observe that the remaining percentage of bookings (almost 60%), made through Online Travel Agents get canceled the maximum number of times.

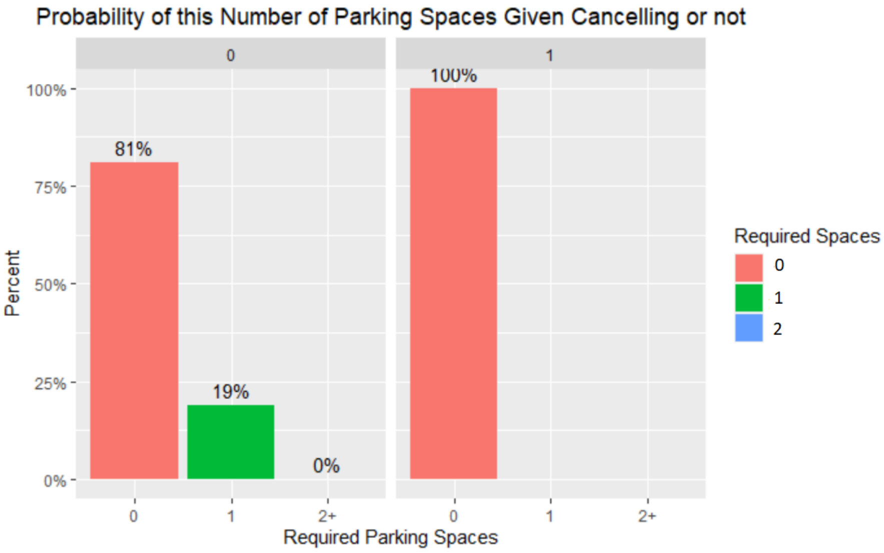


As the maximum bookings cancelled belong to the Online TA mode, we need to check if the bookings made through this mode are legitimate. If the bookings made though Online TAs have any glitch or do not provide customers with the requirements, then we can encourage people to use the other modes of booking which are trustworthy and until then work with the different Online TAs to sort the issue.

13. Parking Spaces Analysis

From the following graph, we have inferred that the 100% of customers tend to cancel their bookings when they are not provided with a parking space (in this case: 0)

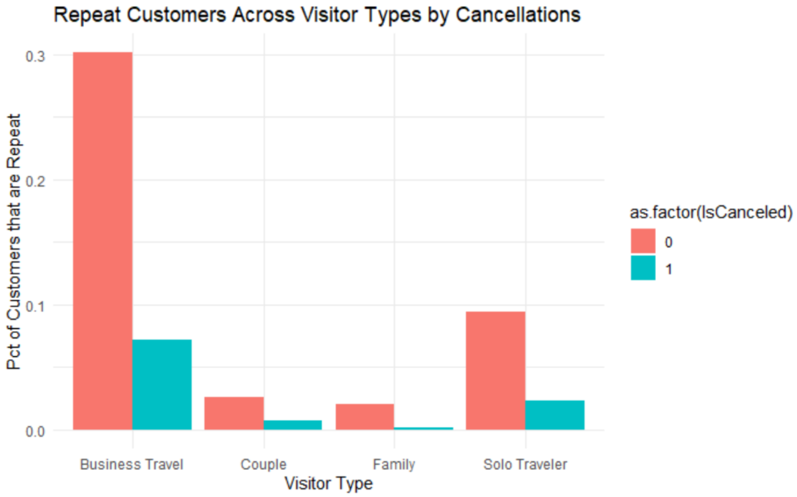
Hence, our recommendation is that the hotel should start by providing parking spaces so as to reduce the number of cancellations.



14. Repeated customers across Visitor types

From this bar plot, we found out that 30% of the customers belonging to business travel repeat the hotel reservation and prefer to come back to that hotel.

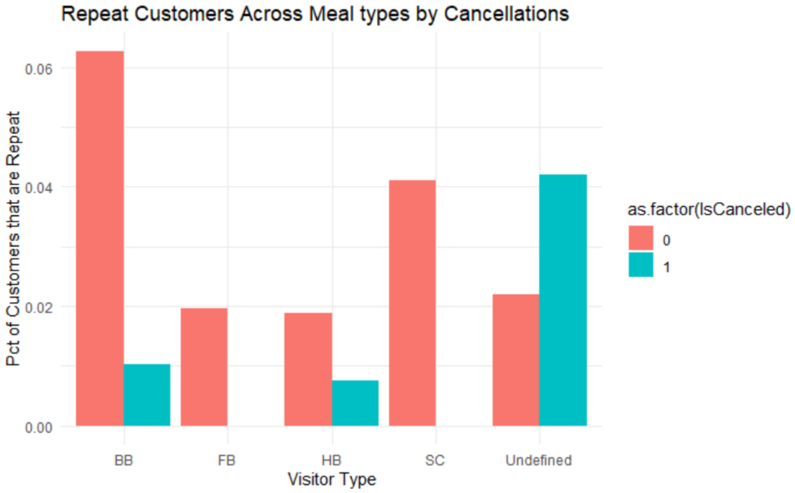
Families and couples seem to not visit the hotel much after the first stay. We can provide families or couples with incentives or offers for the next stay so that they might end up booking our hotel the next time



15. Repeated customers across Meal types

Here we find out that that 6% of the people who have already visited the hotel choose their meal type as bed and breakfast and do not cancel the booking. When the meal type is undefined, 4% of the customers who visit the hotel before cancelling their booking.

So here the recommendation is, if we want to increase the booking rate in the hotel, Bed & Breakfast should be provided.



16. Map Analysis:

Map For Number of Bookings cancelled

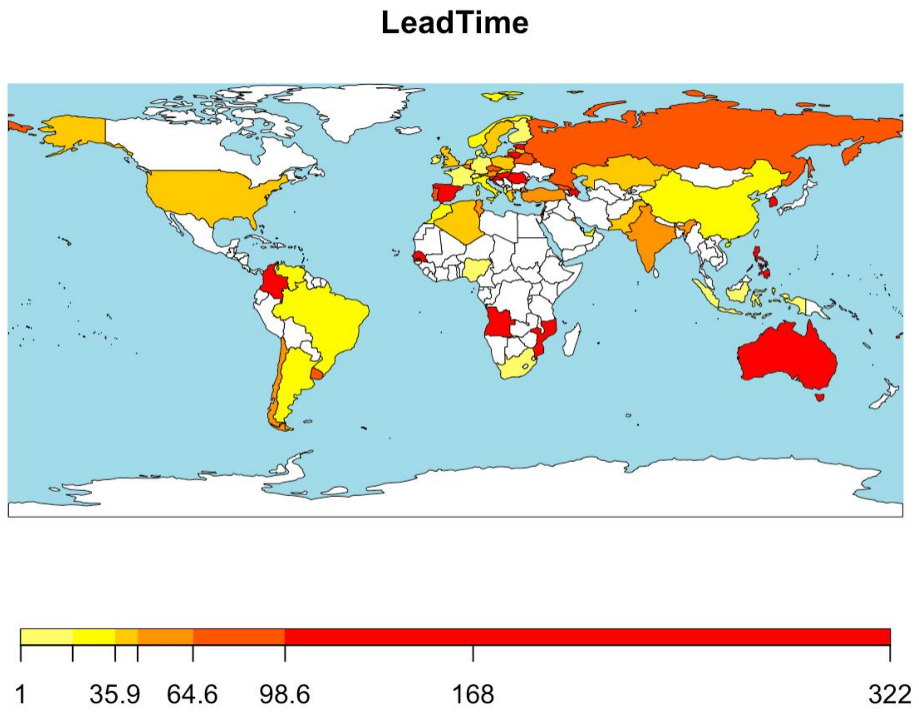
The first map provides us with the countries along with the lead times and we have compared that to the number of bookings cancelled.

The scale provided below the graph depicts the following:

a. Yellow when the lead time is between 1 to 36.
b. The colors subsequently get darker, and the color is the darkest red when the lead time is between 168 to 322.

From the following graph, we can infer that:

a. When the Leadtime is between 168 to 322, the cancellations are maximum. For eg: Australia
b. Bookings are cancelled in China when the Leadtime is between 0 and 36 days.

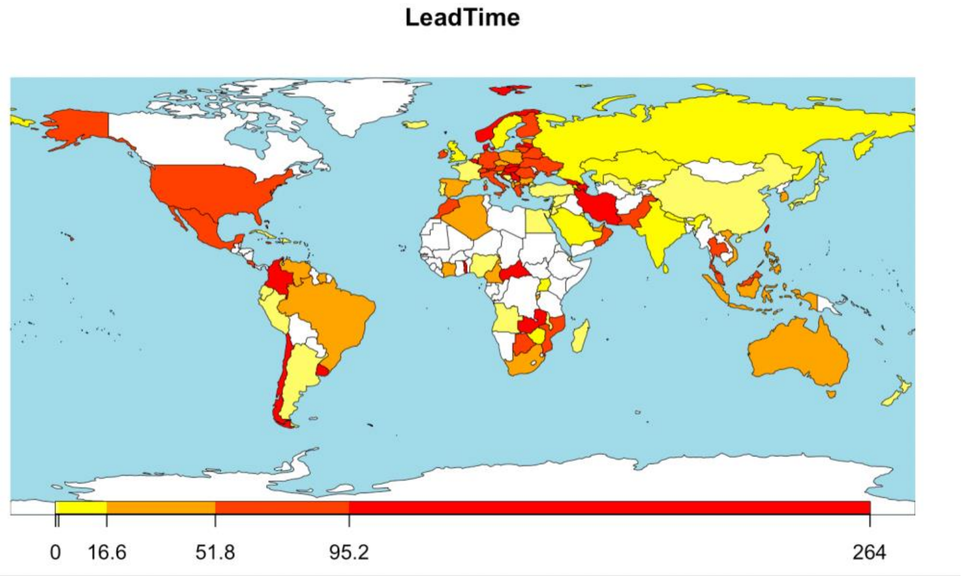


Map for Number of Bookings Not Cancelled

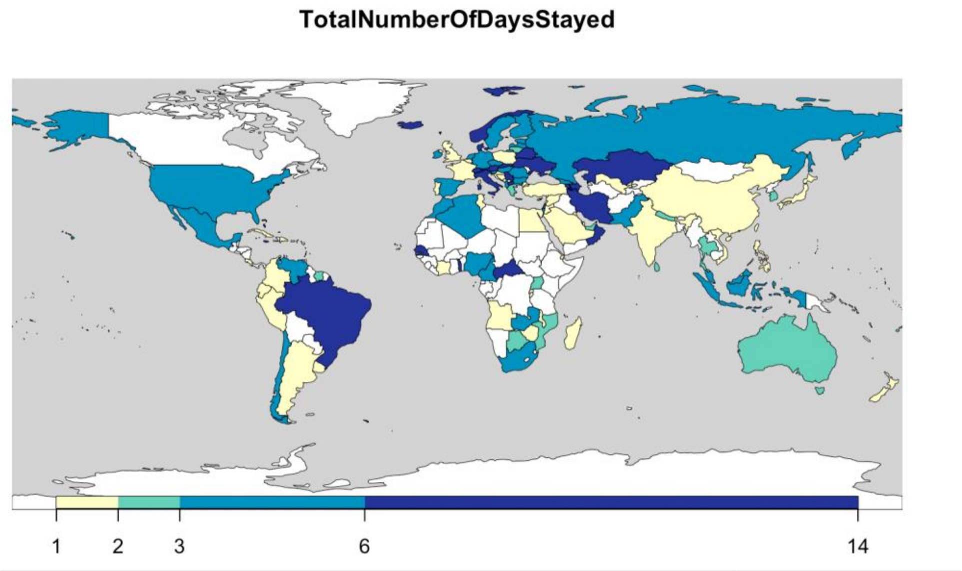
The second map provides us with the countries along with the lead times and we have compared that to the number of bookings that were not cancelled.

The scale provided below the graph depicts the following:

a. Yellow when the lead time is between 0 to 16.6
b. Dark red when the lead time is between 95.2 to 264



Map for Total Number of Days Stayed



The above map depicts the total number of days stayed by the customers from all the countries. The color graph below the map shows the following parameters: Light yellow for 1 to 2 days. Green for 2 to 3 days. Blue for 3 to 6 days. Dark blue for 6 to 14 days (about 2 weeks).

From the above map, we have inferred that customer from countries like Russia, Kazakhstan, etc. stay for the greatest number of days, i.e., between 6-14. On the other hand, customers from countries like India, China, etc. stay for the least number of days i.e., between 1-2.

From this, we recommend that we provide incentives to customers from India, China, etc. So that they can increase their stay duration which in turn will generate revenue and increase the profits of our hotel

Models and Analysis:

1. **Linear and Multiple Regression:**

Following equation was used for the linear model to predict the Cancellation of Hotel Bookings:

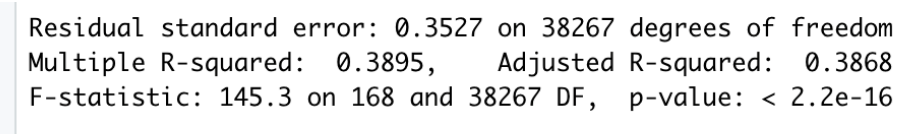
Cancellation Model 1:

The first model we have used is the multiple regression model (lmOut_1), wherein we have predicted IsCanceled using all the variables. After running the summary function on lmOut_1, we have the following inferences:

a. Residual standard error: A measure used to assess how well a linear regression model fits the data. According to the results, our trained model is a 35% fit to our original dataset.
b. Multiple R-squared value: The multiple correlation coefficient between three or more variables. According to our results, our multiple R-squared value is 0.3895 ~ 38.95%
c. Adjusted R-squared value: It is a modified version of R-squared that adjusts for the number of predictors in a regression model. According to our results, the adjusted R-squared value is 0.3868 ~ 38.68%

lmOut_1 <- lm(formula = IsCanceled ~ . , data = df2)

summary(lmOut_1)



```
Residual standard error: 0.3527 on 38267 degrees of freedom
Multiple R-squared:  0.3895,    Adjusted R-squared:  0.3868
F-statistic: 145.3 on 168 and 38267 DF,  p-value: < 2.2e-16
```

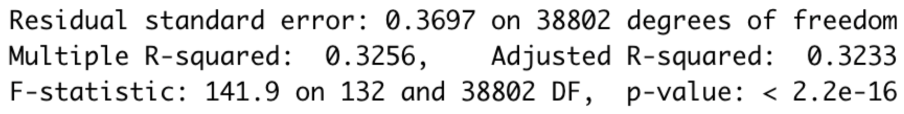
Cancellation Model 2:

The second model we have used is also the multiple regression model (lmOut_2), wherein we have predicted IsCanceled using all the 5 variables. This is because while predicting the linear model of all the attributes, we found out that these were the top six columns having the highest Multiple R-squared values. After running the summary function on lmOut_2, we have the following inferences:

a. Residual standard error: A measure used to assess how well a linear regression model fits the data. According to the results, our trained model is a 36% fit to our original dataset.
b. Multiple R-squared value: The multiple correlation coefficient between three or more variables. According to our results, our multiple R-squared value is 0.3256 ~ 32.56%
c. Adjusted R-squared value: It is a modified version of R-squared that adjusts for the number of predictors in a regression model. According to our results, the adjusted R-squared value is 0.3233 ~ 32.33%

lmOut_2 <- lm(IsCanceled ~ MarketSegment +LeadTime + DepositType + RequiredCarParkingSpaces + Country, data = df2)

summary(lmOut_2)



```
Residual standard error: 0.3697 on 38802 degrees of freedom
Multiple R-squared:  0.3256,    Adjusted R-squared:  0.3233
F-statistic: 141.9 on 132 and 38802 DF,  p-value: < 2.2e-16
```

2. SVM Model:

A support vector machine (SVM) is a machine learning algorithm that analyzes data for classification and regression analysis.

In our dataset, we have used SVM model to find the accuracy for our trained model. We have split the dataset into 60% training dataset and 40% testing dataset.

Two important parameters are used in SVM called Cost and Cross validation. We have kept the cost=3 and cross validation=5 in the model.

The SVM model has a cost function, which controls training errors and margins. For example, a small cost creates a large margin (a soft margin) and allows more misclassifications. On the other hand, a large cost creates a narrow margin (a hard margin) and permits fewer misclassifications.

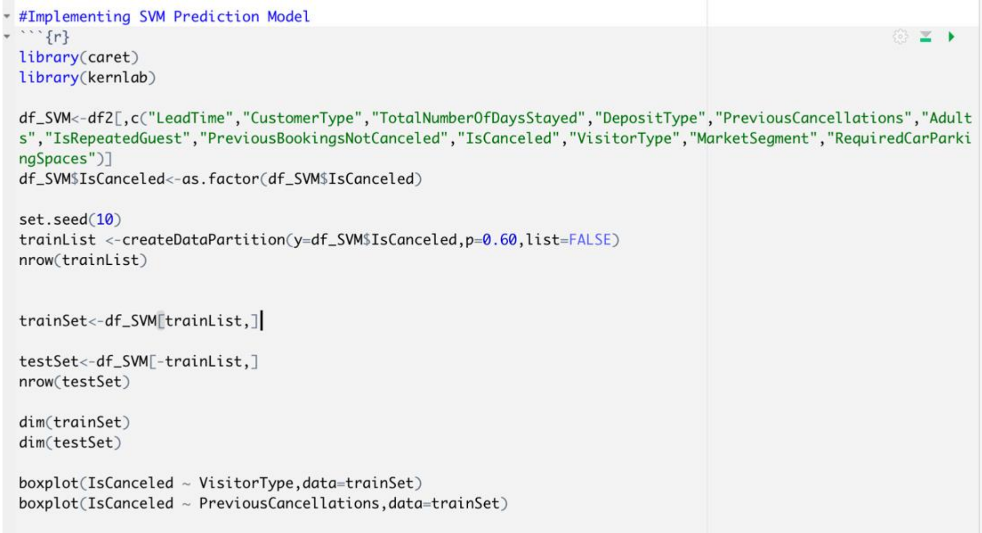
Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations.

First, we implemented SVM on all the 21 columns in the dataset that were present. The accuracy of that model was 76%.

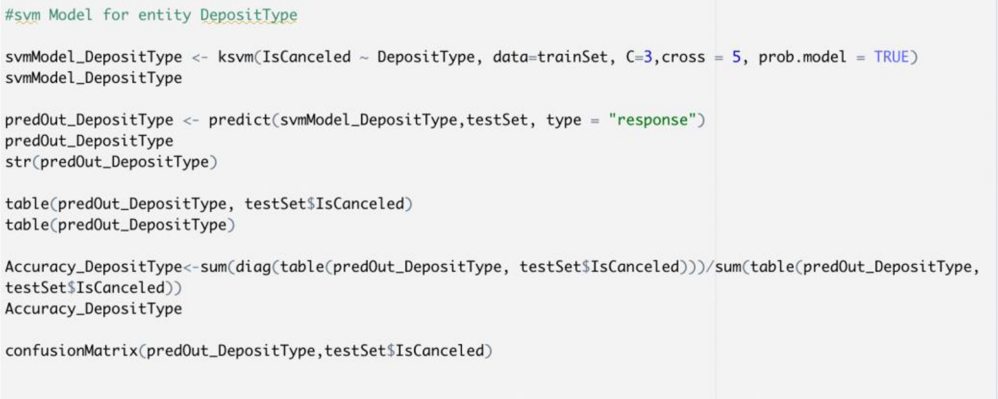
Then, we tried implementing the model for Deposit type, Customer Type, Market Segment, required parking spaces, previous booking cancellations and lead time as we felt these were the most significant attributes from the linear model. We observed that the accuracy for that model was 79%.

After that, we created a new column Visitor type in which the visitor types were combined into 4 categories like family, business travel, solo traveler and couples. When we trained the model on these categories, we found out that the accuracy had further increased to 80%. This is the best accuracy one can achieve from the model.

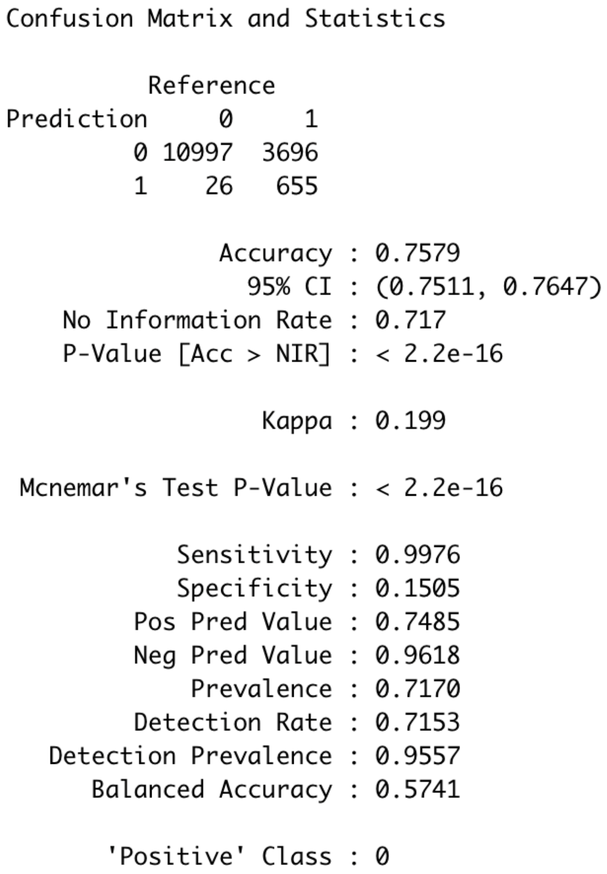
We use confusion matrix to find out the accuracy of the model. A confusion matrix in R is a table that will categorize the predictions against the actual values. It includes two dimensions; among them one will indicate the predicted values and another one will represent the actual values.



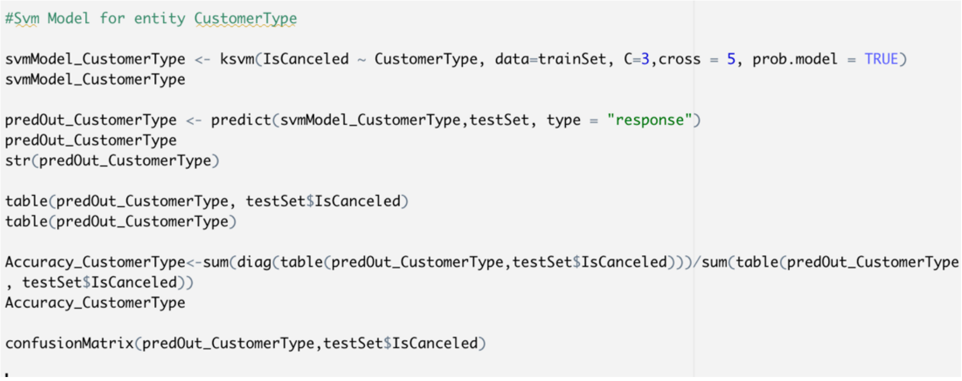
~~~
#Implementing SVM Prediction Model
```{r}
library(caret)
library(kernlab)

df_SVM<-df2[,c("LeadTime","CustomerType","TotalNumberOfDaysStayed","DepositType","PreviousCancellations","Adult
s","IsRepeatedGuest","PreviousBookingsNotCanceled","IsCanceled","VisitorType","MarketSegment","RequiredCarParki
ngSpaces")]
df_SVM$IsCanceled<-as.factor(df_SVM$IsCanceled)

set.seed(10)
trainList <-createDataPartition(y=df_SVM$IsCanceled,p=0.60,list=FALSE)
nrow(trainList)


trainSet<-df_SVM[trainList,]

testSet<-df_SVM[-trainList,]
nrow(testSet)

dim(trainSet)
dim(testSet)

boxplot(IsCanceled ~ VisitorType,data=trainSet)
boxplot(IsCanceled ~ PreviousCancellations,data=trainSet)
~~~



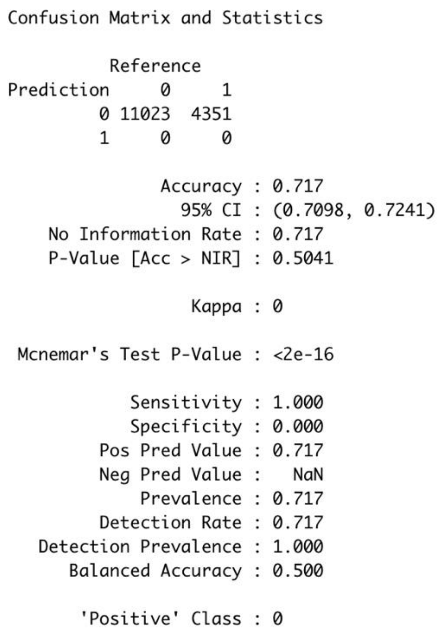
```
#svm Model for entity DepositType

svmModel_DepositType <- ksvm(IsCanceled ~ DepositType, data=trainSet, C=3,cross = 5, prob.model = TRUE)
svmModel_DepositType

predOut_DepositType <- predict(svmModel_DepositType,testSet, type = "response")
predOut_DepositType
str(predOut_DepositType)

table(predOut_DepositType, testSet$IsCanceled)
table(predOut_DepositType)

Accuracy_DepositType<-sum(diag(table(predOut_DepositType, testSet$IsCanceled)))/sum(table(predOut_DepositType,
testSet$IsCanceled))
Accuracy_DepositType

confusionMatrix(predOut_DepositType,testSet$IsCanceled)
```



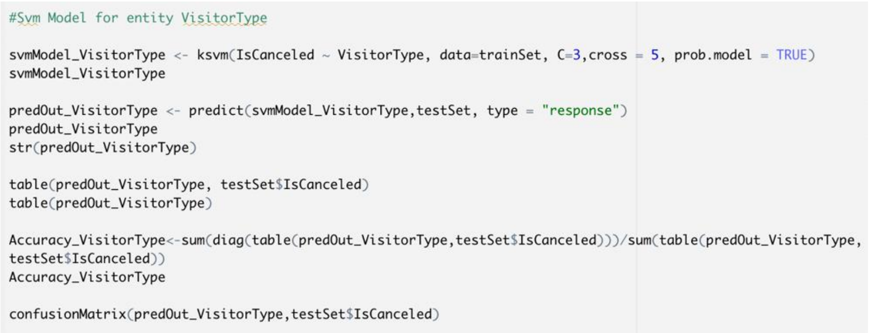
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10997  3696
         1    26   655
                                          
               Accuracy : 0.7579          
                 95% CI : (0.7511, 0.7647)
    No Information Rate : 0.717           
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.199           
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.9976          
            Specificity : 0.1505          
         Pos Pred Value : 0.7485          
         Neg Pred Value : 0.9618          
             Prevalence : 0.7170          
         Detection Rate : 0.7153          
   Detection Prevalence : 0.9557          
      Balanced Accuracy : 0.5741          
                                          
       'Positive' Class : 0               
```



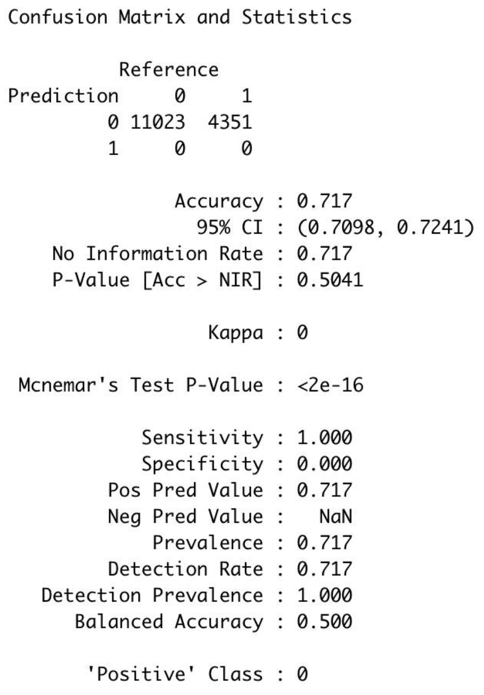
```
#Svm Model for entity CustomerType

svmModel_CustomerType <- ksvm(IsCanceled ~ CustomerType, data=trainSet, C=3,cross = 5, prob.model = TRUE)
svmModel_CustomerType

predOut_CustomerType <- predict(svmModel_CustomerType,testSet, type = "response")
predOut_CustomerType
str(predOut_CustomerType)

table(predOut_CustomerType, testSet$IsCanceled)
table(predOut_CustomerType)

Accuracy_CustomerType<-sum(diag(table(predOut_CustomerType,testSet$IsCanceled)))/sum(table(predOut_CustomerType
, testSet$IsCanceled))
Accuracy_CustomerType

confusionMatrix(predOut_CustomerType,testSet$IsCanceled)
```



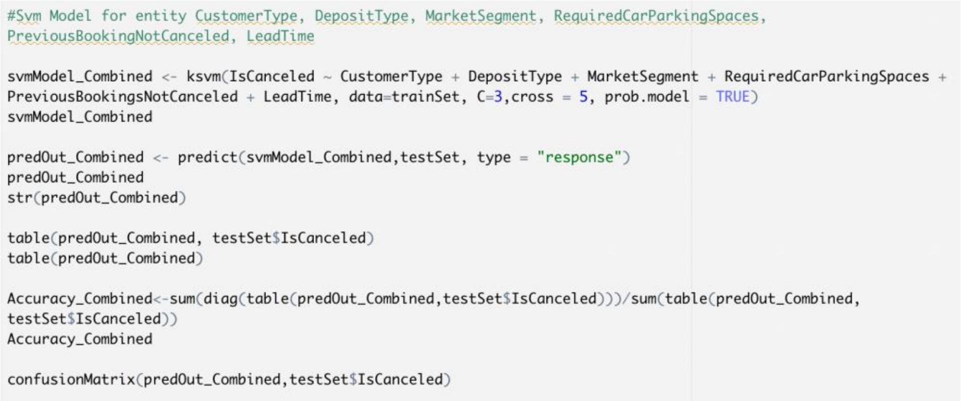
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11023  4351
         1     0     0
                                          
               Accuracy : 0.717           
                 95% CI : (0.7098, 0.7241)
    No Information Rate : 0.717           
    P-Value [Acc > NIR] : 0.5041          
                                          
                  Kappa : 0               
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 1.000           
            Specificity : 0.000           
         Pos Pred Value : 0.717           
         Neg Pred Value :   NaN           
             Prevalence : 0.717           
         Detection Rate : 0.717           
   Detection Prevalence : 1.000           
      Balanced Accuracy : 0.500           
                                          
       'Positive' Class : 0               
```



```
#Svm Model for entity VisitorType

svmModel_VisitorType <- ksvm(IsCanceled ~ VisitorType, data=trainSet, C=3,cross = 5, prob.model = TRUE)
svmModel_VisitorType

predOut_VisitorType <- predict(svmModel_VisitorType,testSet, type = "response")
predOut_VisitorType
str(predOut_VisitorType)

table(predOut_VisitorType, testSet$IsCanceled)
table(predOut_VisitorType)

Accuracy_VisitorType<-sum(diag(table(predOut_VisitorType,testSet$IsCanceled)))/sum(table(predOut_VisitorType,
testSet$IsCanceled))
Accuracy_VisitorType

confusionMatrix(predOut_VisitorType,testSet$IsCanceled)
```

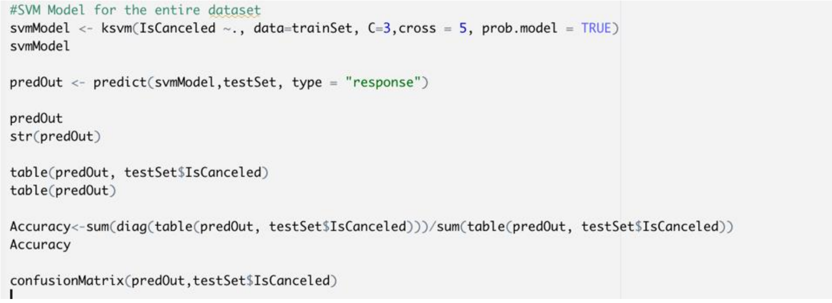


```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11023  4351
         1     0     0
                                          
               Accuracy : 0.717           
                 95% CI : (0.7098, 0.7241)
    No Information Rate : 0.717           
    P-Value [Acc > NIR] : 0.5041          
                                          
                  Kappa : 0               
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 1.000           
            Specificity : 0.000           
         Pos Pred Value : 0.717           
         Neg Pred Value :   NaN           
             Prevalence : 0.717           
         Detection Rate : 0.717           
   Detection Prevalence : 1.000           
      Balanced Accuracy : 0.500           
                                          
       'Positive' Class : 0               
```

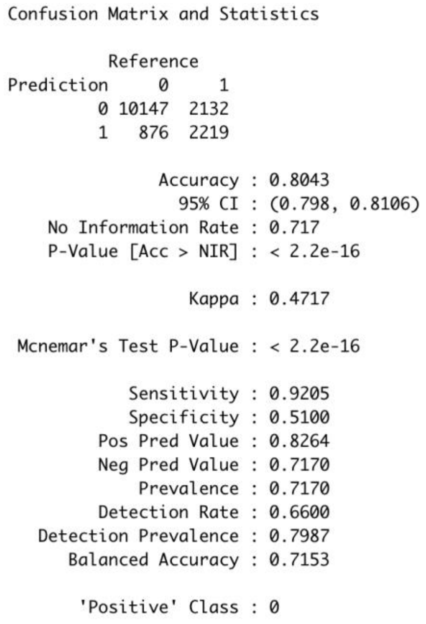


```
#Svm Model for entity CustomerType, DepositType, MarketSegment, RequiredCarParkingSpaces,
PreviousBookingNotCanceled, LeadTime

svmModel_Combined <- ksvm(IsCanceled ~ CustomerType + DepositType + MarketSegment + RequiredCarParkingSpaces +
PreviousBookingsNotCanceled + LeadTime, data=trainSet, C=3,cross = 5, prob.model = TRUE)
svmModel_Combined

predOut_Combined <- predict(svmModel_Combined,testSet, type = "response")
predOut_Combined
str(predOut_Combined)

table(predOut_Combined, testSet$IsCanceled)
table(predOut_Combined)

Accuracy_Combined<-sum(diag(table(predOut_Combined,testSet$IsCanceled)))/sum(table(predOut_Combined,
testSet$IsCanceled))
Accuracy_Combined

confusionMatrix(predOut_Combined,testSet$IsCanceled)
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10030  2207
         1   993  2144
                                          
               Accuracy : 0.7919          
                 95% CI : (0.7854, 0.7983)
    No Information Rate : 0.717           
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.4398          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.9099          
            Specificity : 0.4928          
         Pos Pred Value : 0.8196          
         Neg Pred Value : 0.6835          
             Prevalence : 0.7170          
         Detection Rate : 0.6524          
   Detection Prevalence : 0.7960          
      Balanced Accuracy : 0.7013          
                                          
       'Positive' Class : 0               
```

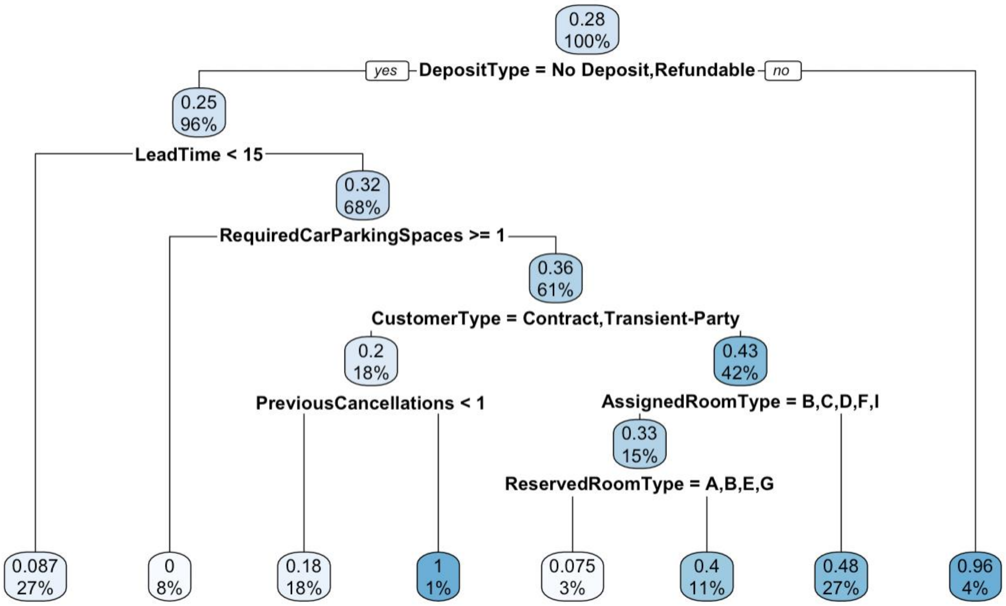


```
#SVM Model for the entire dataset
svmModel <- ksvm(IsCanceled ~., data=trainSet, C=3,cross = 5, prob.model = TRUE)
svmModel

predOut <- predict(svmModel,testSet, type = "response")

predOut
str(predOut)

table(predOut, testSet$IsCanceled)
table(predOut)

Accuracy<-sum(diag(table(predOut, testSet$IsCanceled)))/sum(table(predOut, testSet$IsCanceled))
Accuracy

confusionMatrix(predOut,testSet$IsCanceled)
```



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10147  2132
         1   876  2219

               Accuracy : 0.8043
                 95% CI : (0.798, 0.8106)
    No Information Rate : 0.717
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4717

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

            Sensitivity : 0.9205
            Specificity : 0.5100
         Pos Pred Value : 0.8264
         Neg Pred Value : 0.7170
             Prevalence : 0.7170
         Detection Rate : 0.6600
   Detection Prevalence : 0.7987
      Balanced Accuracy : 0.7153

       'Positive' Class : 0
```

3. RPart Binary Tree:


0.28
100%
yes
DepositType = No Deposit,Refundable
no
0.25
96%
LeadTime < 15
0.32
68%
RequiredCarParkingSpaces >= 1
0.36
61%
CustomerType = Contract,Transient-Party
0.2
18%
PreviousCancellations < 1
0.43
42%
AssignedRoomType = B,C,D,F,I
0.33
15%
ReservedRoomType = A,B,E,G
0.087
27%
0
8%
0.18
18%
1
1%
0.075
3%
0.4
11%
0.48
27%
0.96
4%

R’s rpart package provides a powerful framework for growing classification and regression trees. Rpart stands for recursive partitioning.

We make sure all the categorical variables are converted into factors.

In this the main aim is to check whether a booking is cancelled or not. So,:

the top node gives us the probability of the booking being canceled which is 28%. So, it means that 28% of the time the booking will be cancelled which is the same as the one we found while taking the mean of the IsCanceled column.

Deposit type becomes our second criteria for classifying the tree.

The left daughter is the time when IsCanceled=1 and the right daughter is the time when IsCanceled =0.

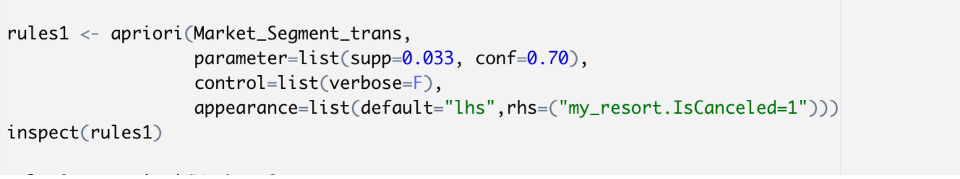
When the deposit type is No Deposit or refundable, we see that 25% people cancel their booking. When the deposit type is not no deposit or not refundable, that means 96% of the people will not cancel their booking.

4. Apriori algorithm and association rules:

Apriori algorithm is used for finding frequent itemsets in a dataset for association rule mining. It is called Apriori because it uses prior knowledge of frequent itemset properties.

The following logic was used to derive Apriori algorithm rules:

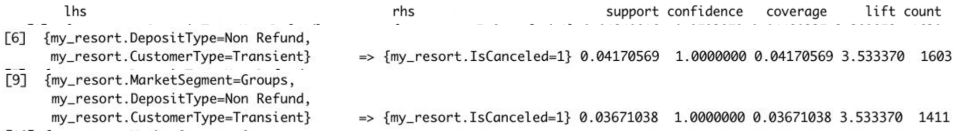
1. The data frame was converted into a transactional matrix
2. Following code was run to derive rules for hotel bookings that were about to cancel



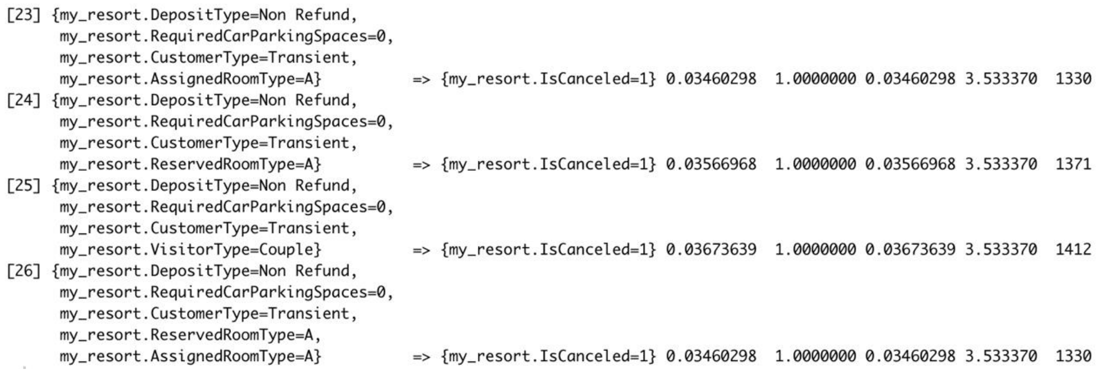
```
rules1 <- apriori(Market_Segment_trans,
                  parameter=list(supp=0.033, conf=0.70),
                  control=list(verbose=F),
                  appearance=list(default="lhs",rhs=("my_resort.IsCanceled=1")))
inspect(rules1)
```

Here the support factor is 0.033 and the confidence factor is 0.70.
The RHS was set as IsCanceled=1 i.e we are predicting whether the bookings are going to canceled and under what conditions
A total of 26 rules were generated out of which 12 rules had confidence as 1.

Rules:



```
      lhs                                          rhs                          support confidence   coverage     lift count
[6]   {my_resort.DepositType=Non Refund,
       my_resort.CustomerType=Transient}       => {my_resort.IsCanceled=1} 0.04170569   1.0000000 0.04170569 3.533370   1603
[9]   {my_resort.MarketSegment=Groups,
       my_resort.DepositType=Non Refund,
       my_resort.CustomerType=Transient}       => {my_resort.IsCanceled=1} 0.03671038   1.0000000 0.03671038 3.533370   1411
```

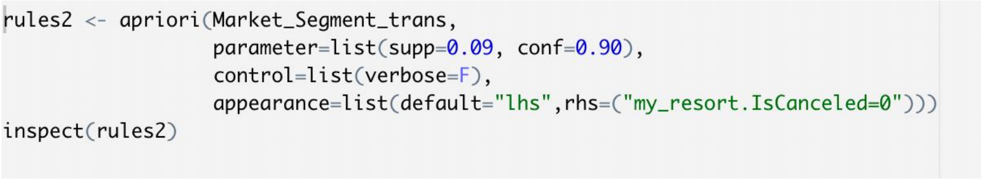


```
[23] {my_resort.DepositType=Non Refund,
      my_resort.RequiredCarParkingSpaces=0,
      my_resort.CustomerType=Transient,
      my_resort.AssignedRoomType=A}         => {my_resort.IsCanceled=1} 0.03460298  1.0000000 0.03460298 3.533370  1330
[24] {my_resort.DepositType=Non Refund,
      my_resort.RequiredCarParkingSpaces=0,
      my_resort.CustomerType=Transient,
      my_resort.ReservedRoomType=A}         => {my_resort.IsCanceled=1} 0.03566968  1.0000000 0.03566968 3.533370  1371
[25] {my_resort.DepositType=Non Refund,
      my_resort.RequiredCarParkingSpaces=0,
      my_resort.CustomerType=Transient,
      my_resort.VisitorType=Couple}         => {my_resort.IsCanceled=1} 0.03673639  1.0000000 0.03673639 3.533370  1412
[26] {my_resort.DepositType=Non Refund,
      my_resort.RequiredCarParkingSpaces=0,
      my_resort.CustomerType=Transient,
      my_resort.ReservedRoomType=A,
      my_resort.AssignedRoomType=A}         => {my_resort.IsCanceled=1} 0.03460298  1.0000000 0.03460298 3.533370  1330
```

Rule 26 says that when the Deposit Type = Non-Refund, Required Car Parking Spaces =0, Customer Type is Transient, Reserved Room Type is A and Assigned Room Type is A, the booking is going to be cancelled. This is because the rule has confidence as 1 and support as 0.034.

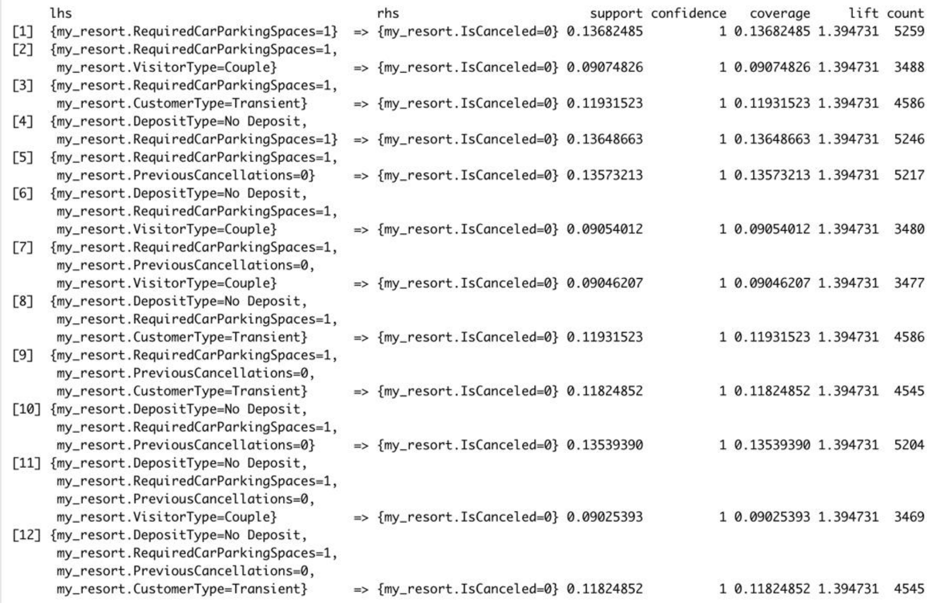
In the similar manner, 25 other rules are developed for predicting under which conditions the bookings might get cancelled. All the bookings have a support level above 0.033 and confidence level above 0.70.

3. Following code was run to derive rules for hotel bookings that won’t be cancelled.



```
rules2 <- apriori(Market_Segment_trans,
                  parameter=list(supp=0.09, conf=0.90),
                  control=list(verbose=F),
                  appearance=list(default="lhs",rhs=("my_resort.IsCanceled=0")))
inspect(rules2)
```

Here the support factor is 0.09 and the confidence factor is 0.90.
The RHS was set as IsCanceled=0 i.e we are predicting whether the bookings won't be canceled and under what conditions
A total of 12 rules were generated out of which all the rules have confidence as 1.



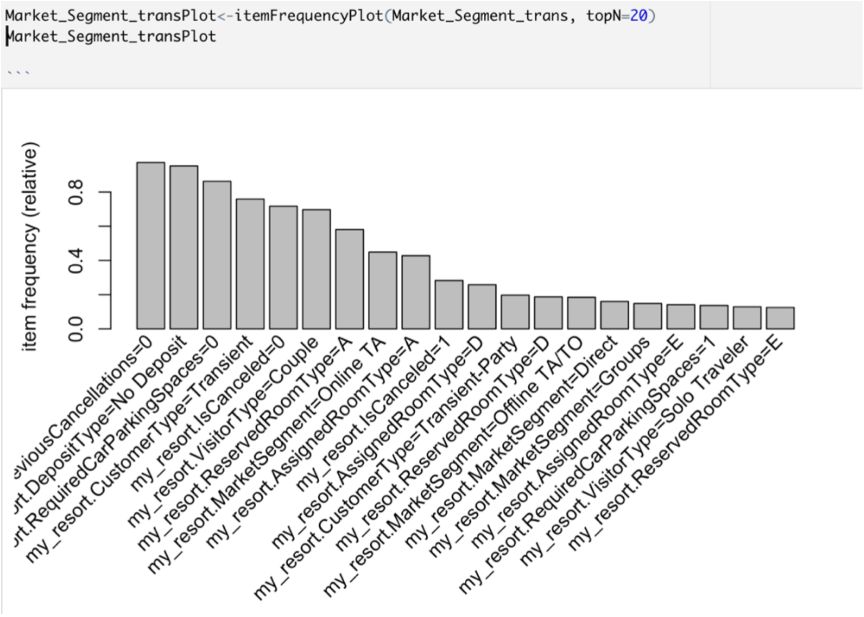
```
     lhs                                         rhs                            support confidence   coverage     lift count
[1]  {my_resort.RequiredCarParkingSpaces=1}  => {my_resort.IsCanceled=0} 0.13682485          1 0.13682485 1.394731  5259
[2]  {my_resort.RequiredCarParkingSpaces=1,
      my_resort.VisitorType=Couple}          => {my_resort.IsCanceled=0} 0.09074826          1 0.09074826 1.394731  3488
[3]  {my_resort.RequiredCarParkingSpaces=1,
      my_resort.CustomerType=Transient}      => {my_resort.IsCanceled=0} 0.11931523          1 0.11931523 1.394731  4586
[4]  {my_resort.DepositType=No Deposit,
      my_resort.RequiredCarParkingSpaces=1}  => {my_resort.IsCanceled=0} 0.13648663          1 0.13648663 1.394731  5246
[5]  {my_resort.RequiredCarParkingSpaces=1,
      my_resort.PreviousCancellations=0}     => {my_resort.IsCanceled=0} 0.13573213          1 0.13573213 1.394731  5217
[6]  {my_resort.DepositType=No Deposit,
      my_resort.RequiredCarParkingSpaces=1,
      my_resort.VisitorType=Couple}          => {my_resort.IsCanceled=0} 0.09054012          1 0.09054012 1.394731  3480
[7]  {my_resort.RequiredCarParkingSpaces=1,
      my_resort.PreviousCancellations=0,
      my_resort.VisitorType=Couple}          => {my_resort.IsCanceled=0} 0.09046207          1 0.09046207 1.394731  3477
[8]  {my_resort.DepositType=No Deposit,
      my_resort.RequiredCarParkingSpaces=1,
      my_resort.CustomerType=Transient}      => {my_resort.IsCanceled=0} 0.11931523          1 0.11931523 1.394731  4586
[9]  {my_resort.RequiredCarParkingSpaces=1,
      my_resort.PreviousCancellations=0,
      my_resort.CustomerType=Transient}      => {my_resort.IsCanceled=0} 0.11824852          1 0.11824852 1.394731  4545
[10] {my_resort.DepositType=No Deposit,
      my_resort.RequiredCarParkingSpaces=1,
      my_resort.PreviousCancellations=0}     => {my_resort.IsCanceled=0} 0.13539390          1 0.13539390 1.394731  5204
[11] {my_resort.DepositType=No Deposit,
      my_resort.RequiredCarParkingSpaces=1,
      my_resort.PreviousCancellations=0,
      my_resort.VisitorType=Couple}          => {my_resort.IsCanceled=0} 0.09025393          1 0.09025393 1.394731  3469
[12] {my_resort.DepositType=No Deposit,
      my_resort.RequiredCarParkingSpaces=1,
      my_resort.PreviousCancellations=0,
      my_resort.CustomerType=Transient}      => {my_resort.IsCanceled=0} 0.11824852          1 0.11824852 1.394731  4545
```

The rule number 12 tells that if the Deposit Type is No Deposit, Required Car Parking Spaces is 1 Previous Cancellations are 0 and Customer Type is Transient, the booking won't get cancelled.

This is because the rule has the support of 0.11 and confidence level as 1.

In a similar manner, 11 other rules are developed for predicting under which conditions the bookings won't be cancelled. All the bookings have a support level above 0.09 and confidence level above 0.90.

4. We got the following frequency plot for top 20 items which have highest frequency



```
Market_Segment_transPlot<-itemFrequencyPlot(Market_Segment_trans, topN=20)
Market_Segment_transPlot

```
```
```

Recommendations:

- Our observations have suggested that customers cancel their bookings if they do not get room of their choice. Hence, we recommend that the hotel increases the availability of type A rooms so that more customers will be satisfied and in turn attract more customers.
- We have also observed that customers are more likely to not cancel their bookings if they get a parking spot.
- Customers are also likely to cancel their bookings if they haven't made any previous deposits to confirm their bookings. Hence, we recommend that the hotels should encourage but not force the customers to pay some sort of deposit so that they feel compelled to stick to their bookings. If the customers are not willing to pay the deposit, then a fixed percentage of cancellation fee should be imposed on the customers.
- The hotel should also overbook the rooms for guests specifically from countries like Indonesia or keep the rooms open for last-minute bookings.
- We also recommend that the hotel should take part in corporate programs because we have seen that the corporate bookings have a low rate of cancellation.
- We have observed that the almost 60% of the bookings made through Online TA (Travel Agents) get cancelled. Therefore, we recommend that the hotel rectifies and fixes the problems if there are any within the system. We also recommend that the hotel provides more offers/ rewards for each booking made through an online TA.
- We also recommend that the hotel provides some sort of complementary meals such as 'Bed in breakfast' meals so that customers feel attracted towards such offers and this will lead to more bookings.