IST 687

Data Science Project Report

Hotel Insights, Findings & Recommendations

Professor: Erik Anderson, Jeffrey Saltz

Submitted by:
David Garcia
Hrishikesh Telang
Liam Hogan
Priyal Bhamnikar
Siddhi Bhandari

DESCRIPTION

- The Hotel Industry, in general, has always been volatile when it comes to reservation and cancellation of rooms. Some customers have reserved hotel rooms months, sometimes even a year before arrival. However, many of these customers potentially cancel bookings due to unforeseen reasons and other factors.
- Several attributes cause this to happen. After analysis, we found several factors that when targeted can significantly reduce cancellations. Thus, as data scientists, on behalf of the hotel industry, we aim to understand the historical data and analyze hotel booking trends. These factors can be used to report the trends and predict future bookings.
- With the various insights and analyses obtained, we aim to generate insights that can help drive business decisions on how to reduce cancellation bookings and essentially improve profits in the hotel business.

TECHNICAL DETAILS

- There are 20 features and 40,060 observations available in the dataset.
- The dataset is not overwritten, and columns are not renamed for ease of understanding and coding.
- Upon checking the NULL and missing values, we found out that there are no missing values; however, the NULL values were present in the country column as string values, which we converted into actual null type variables.
- We also identified the categorical values (by converting them into factor variables) and numerical data. We created separate datasets, one to perform association rules mining and another to perform ML classification.

GOALS/ OBJECTIVES

The overall goal of the project is to provide actionable recommendations, based on the insights to prevent or stop the increment of cancellations.

OBJECTIVE

1. To perform analysis on columns such as cancellation, customer type, market segment, among other attributes.
2. Implement various Machine Learning algorithms to predict cancellation.
3. To analyze why people cancel hotel reservations and predict who will be canceling
4. To analyze:
    a) The number of cancellations:
        i) Number of bookings on a weekday vs weekends
        ii) Most preferred meal types
        iii) Country-wise bookings

iv) New customers acquired
v) Type of rooms preferred by customers
vi) Booking types
vii) Assigned Rooms
viii) The number of guests in each booking

b) Analyze patterns associated with each segment, such as:
   i) Day of week
   ii) Type of customers
   iii) Type of rooms
   iv) Market Segment

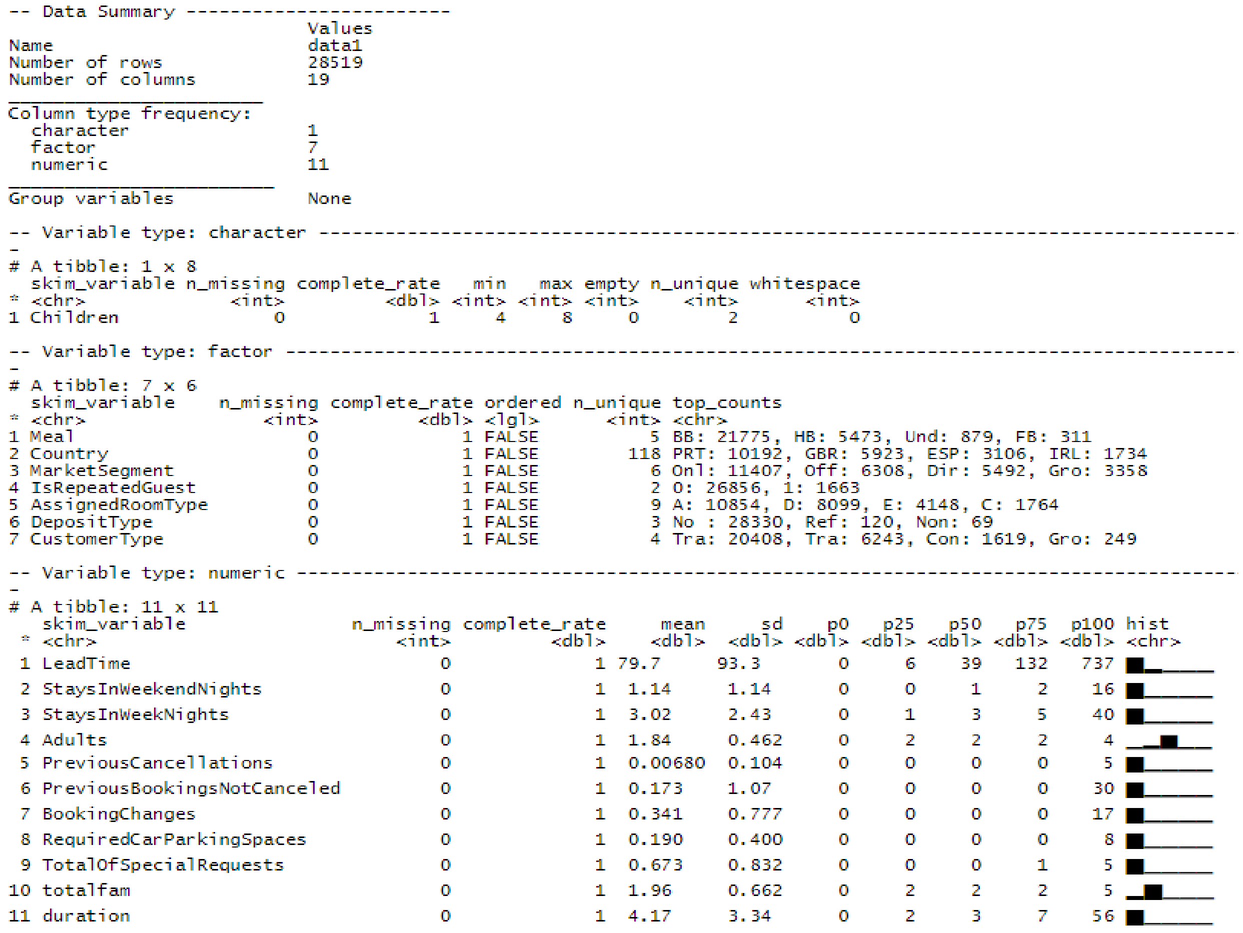
c) Predict future cancellations based on machine learning algorithms such as the Apriori algorithm, linear modeling, support vector machines, and classification and regression trees.
d) Using these results, we can make critical business decisions regarding the customer experience they desire to deliver.

LIBRARIES USED

We used the following libraries for the project
Tidyverse, caret, rworldmap, skimr, ggplot2, arules, readr, rpart, e1071, rpart.plot

DATASET

Upon exploring the structure of the data frame, we notice some variables are characters. Then, we replace those variables as factors to better analyze them.



```
-- Data Summary ------------------------
                           Values
Name                       data1
Number of rows             28519
Number of columns          19
_______________________
Column type frequency:
  character                1
  factor                   7
  numeric                  11
________________________
Group variables            None

-- Variable type: character ---------------------------------------------------------------------------
-
# A tibble: 1 x 8
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
* <chr>             <int>         <dbl> <int> <int> <int>    <int>      <int>
1 Children              0             1     4     8     0        2          0

-- Variable type: factor ------------------------------------------------------------------------------
-
# A tibble: 7 x 6
  skim_variable    n_missing complete_rate ordered n_unique top_counts
* <chr>                <int>         <dbl> <lgl>      <int> <chr>
1 Meal                     0             1 FALSE          5 BB: 21775, HB: 5473, Und: 879, FB: 311
2 Country                  0             1 FALSE        118 PRT: 10192, GBR: 5923, ESP: 3106, IRL: 1734
3 MarketSegment            0             1 FALSE          6 Onl: 11407, Off: 6308, Dir: 5492, Gro: 3358
4 IsRepeatedGuest          0             1 FALSE          2 0: 26856, 1: 1663
5 AssignedRoomType         0             1 FALSE          9 A: 10854, D: 8099, E: 4148, C: 1764
6 DepositType              0             1 FALSE          3 No : 28330, Ref: 120, Non: 69
7 CustomerType             0             1 FALSE          4 Tra: 20408, Tra: 6243, Con: 1619, Gro: 249

-- Variable type: numeric -----------------------------------------------------------------------------
-
# A tibble: 11 x 11
   skim_variable                n_missing complete_rate    mean    sd    p0   p25   p50   p75  p100 hist
 * <chr>                            <int>         <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
 1 LeadTime                             0             1 79.7    93.3      0     6    39   132   737 ▇▂▁▁▁
 2 StaysInWeekendNights                 0             1  1.14    1.14     0     0     1     2    16 ▇▁▁▁▁
 3 StaysInWeekNights                    0             1  3.02    2.43     0     1     3     5    40 ▇▁▁▁▁
 4 Adults                               0             1  1.84    0.462    0     2     2     2     4 ▁▁▇▁▁
 5 PreviousCancellations                0             1  0.00680 0.104    0     0     0     0     5 ▇▁▁▁▁
 6 PreviousBookingsNotCanceled          0             1  0.173   1.07     0     0     0     0    30 ▇▁▁▁▁
 7 BookingChanges                       0             1  0.341   0.777    0     0     0     0    17 ▇▁▁▁▁
 8 RequiredCarParkingSpaces             0             1  0.190   0.400    0     0     0     0     8 ▇▁▁▁▁
 9 TotalOfSpecialRequests               0             1  0.673   0.832    0     0     0     1     5 ▇▁▁▁▁
10 totalfam                             0             1  1.96    0.662    0     2     2     2     5 ▁▇▁▁▁
11 duration                             0             1  4.17    3.34     0     2     3     7    56 ▇▁▁▁▁
```

Fig. 1: Data summary

Missing data:

The below plot displays the missing values to avoid any misinterpretations of data. We did not find missing values, however, NULL values were found and acted upon.

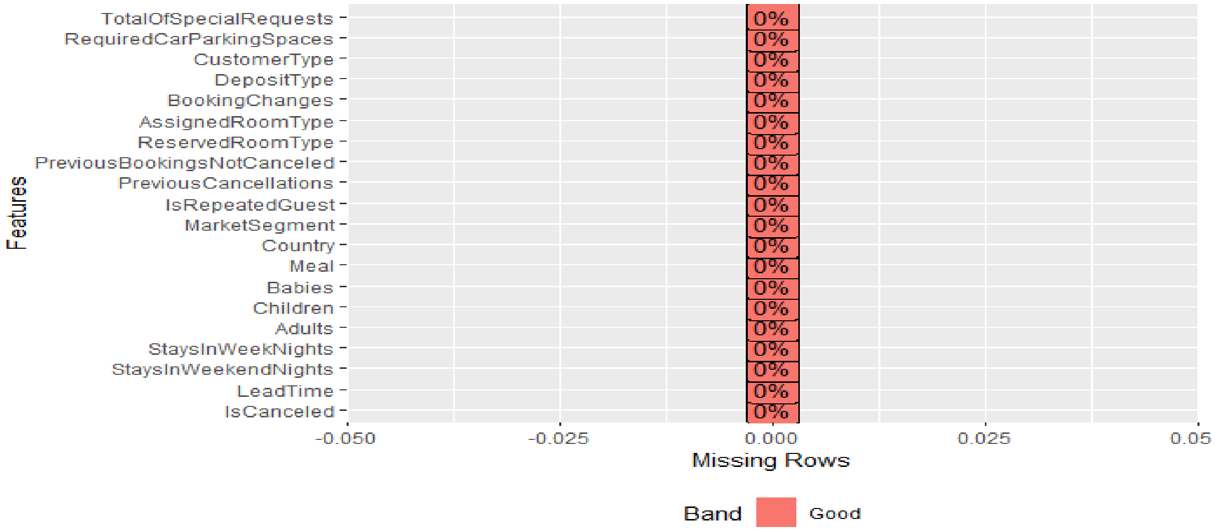


Fig. 2: Missing data percentage

For the exploratory analysis, we created a whole set of visualizations to better understand the data and find some patterns that could be interesting to solve the main question: why are people canceling their reservations?

Let's talk about the cancellations. It is very important to understand why the bookings are canceled in the first place to gain important information on areas of improvement. How many bookings were canceled? What are the factors that affect cancellations?

First, we started with some bar plots to understand the relationship between categorical variables and the categorical variable with cancellation status.

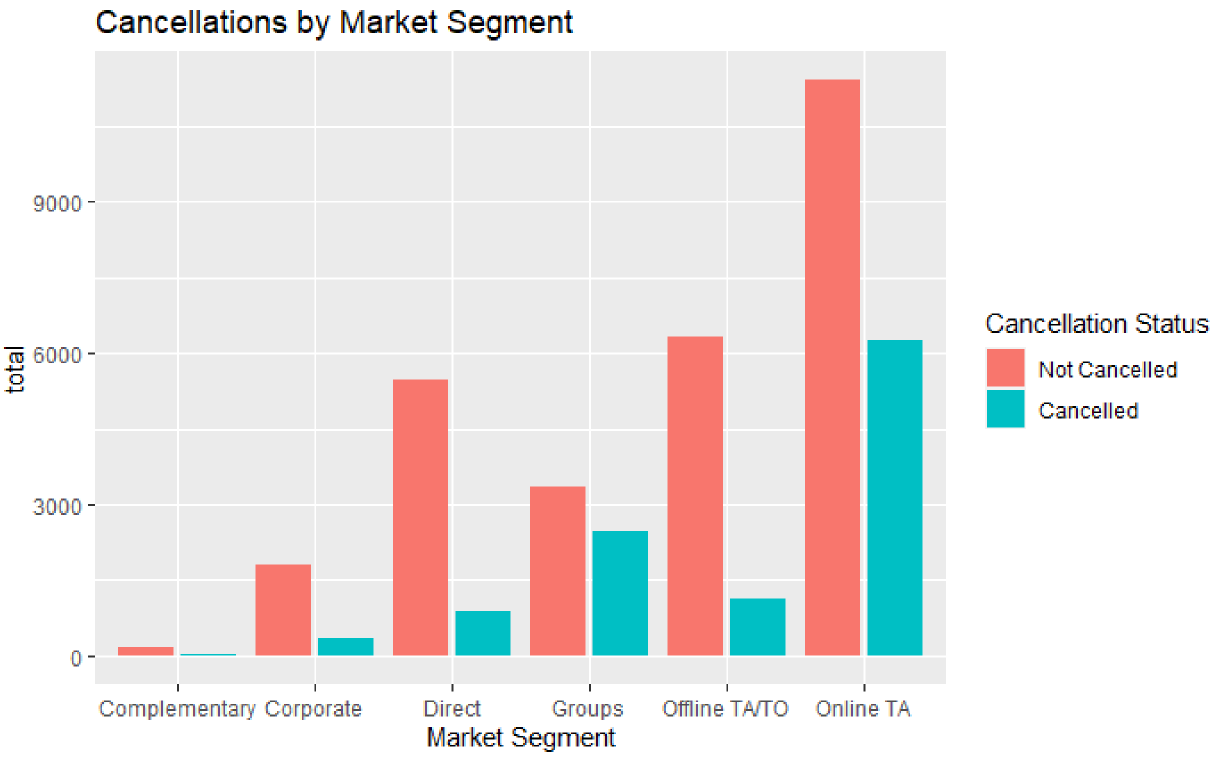


Fig.3 Cancellations by market segment

As we can see in this plot the groups of people that have more cancellations are Online TA and Groups, so later we will focus on market segment and other variables. Moreover, we can see that the hotel is doing correct things with the direct market segment.

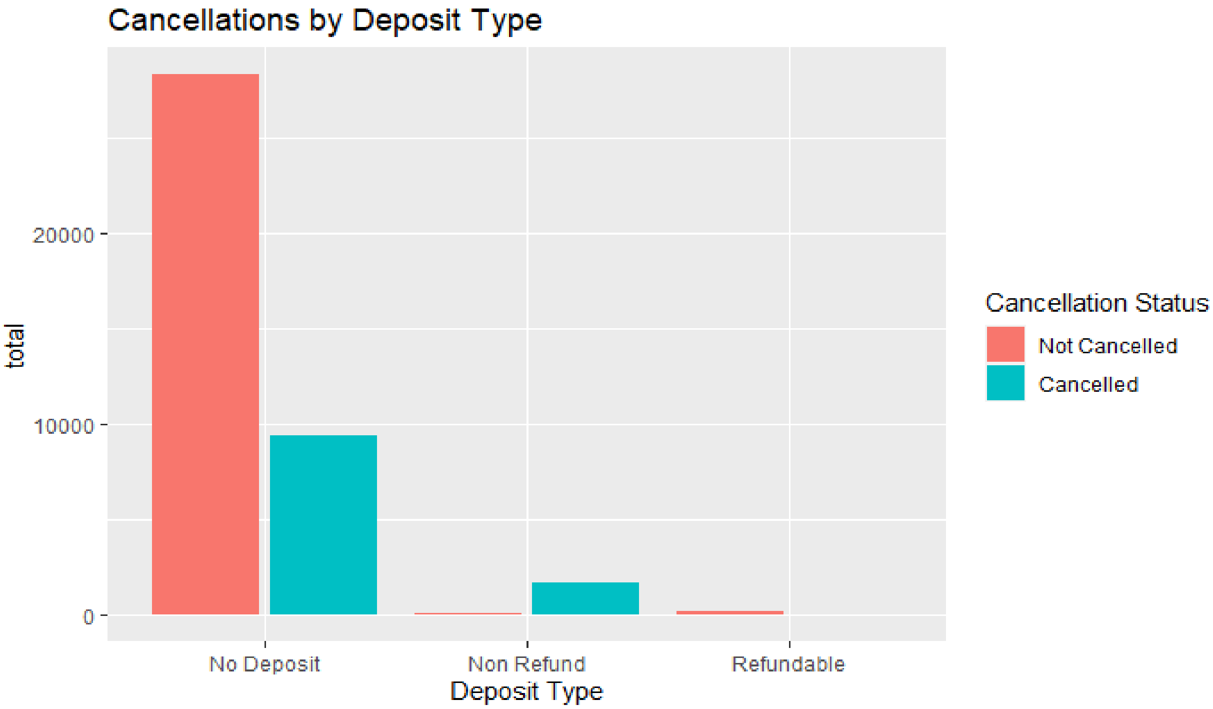


Fig.4 Cancellations by deposit type

The previous chart suggests we should focus on the non-refundable segment of people because their number of cancellations is bigger than the number of no cancellations.

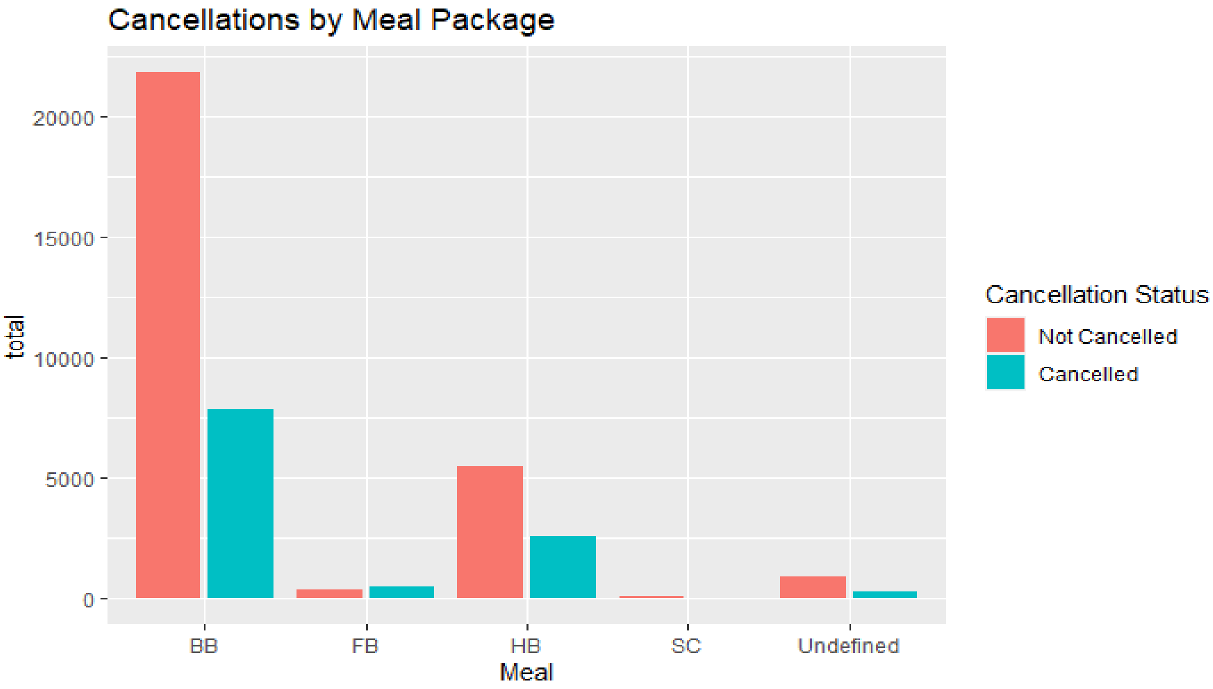


Fig.5 Cancellations by meal package

Related to the meal plan of each person we can see people that took the BB plan canceled more times. At this point we know we should focus on people that have a non-refundable deposit and take the BB meal plan.

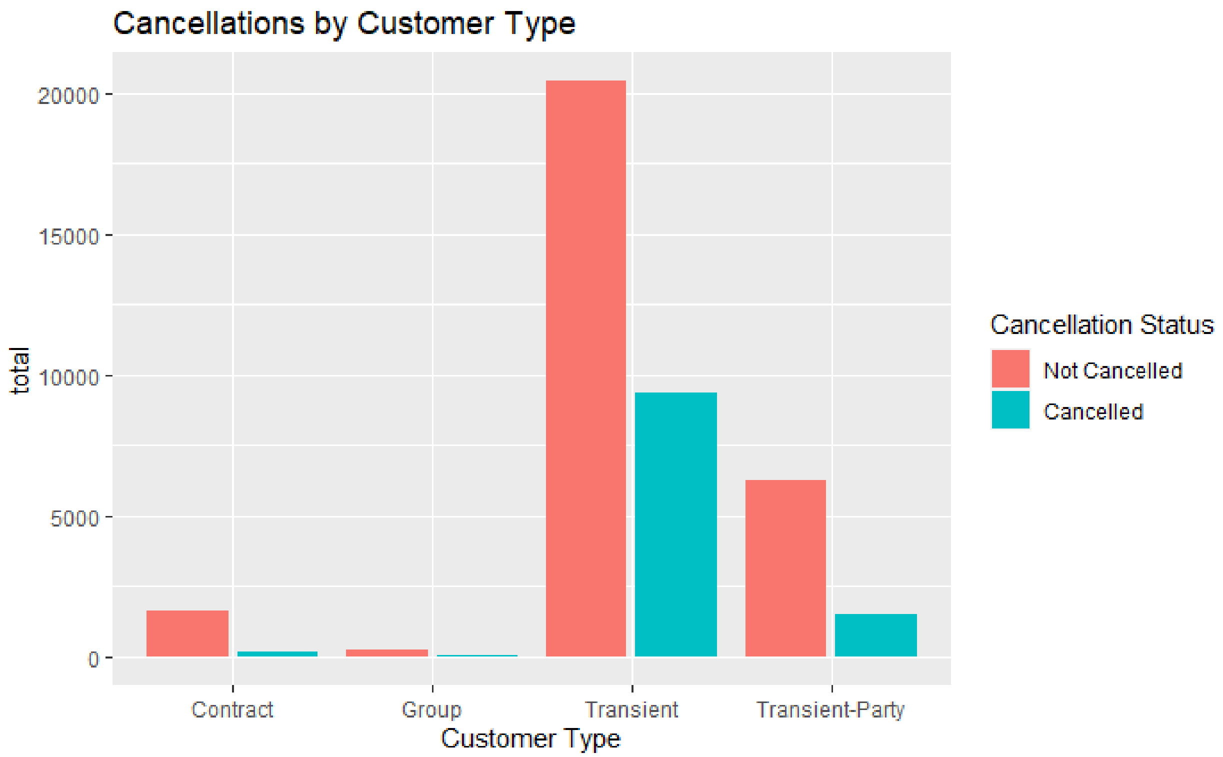


Fig.6 Cancellations by customer type

Another important discover from the data is those people that are considered transient have a large rate of cancellations that is why we should focus on this kind of people.

We also created some boxplots to understand how the numerical variables behave.

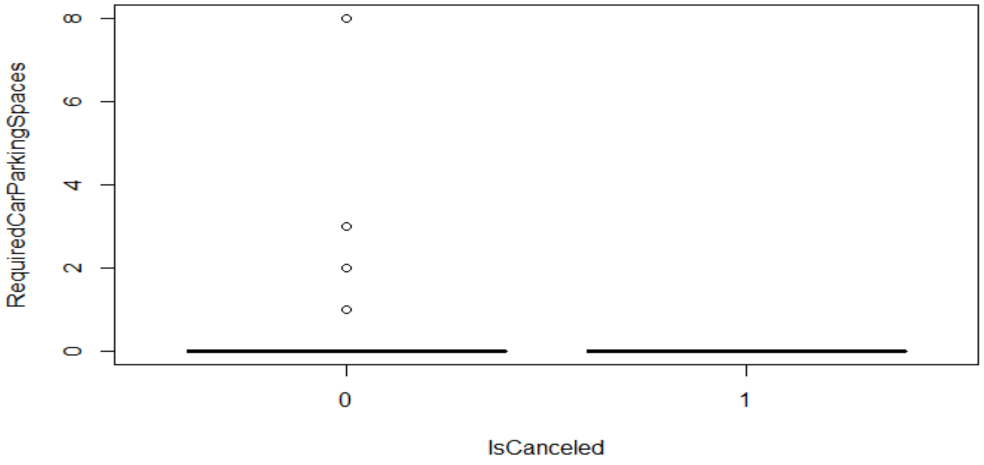


Fig. 7 Boxplot for parking space and cancellations

From this graph, we can see that all the people who canceled didn’t need car spaces at all. This suggests digging deeper into this group of people and probably is correlated with the meal plan and deposit type that we saw previously.

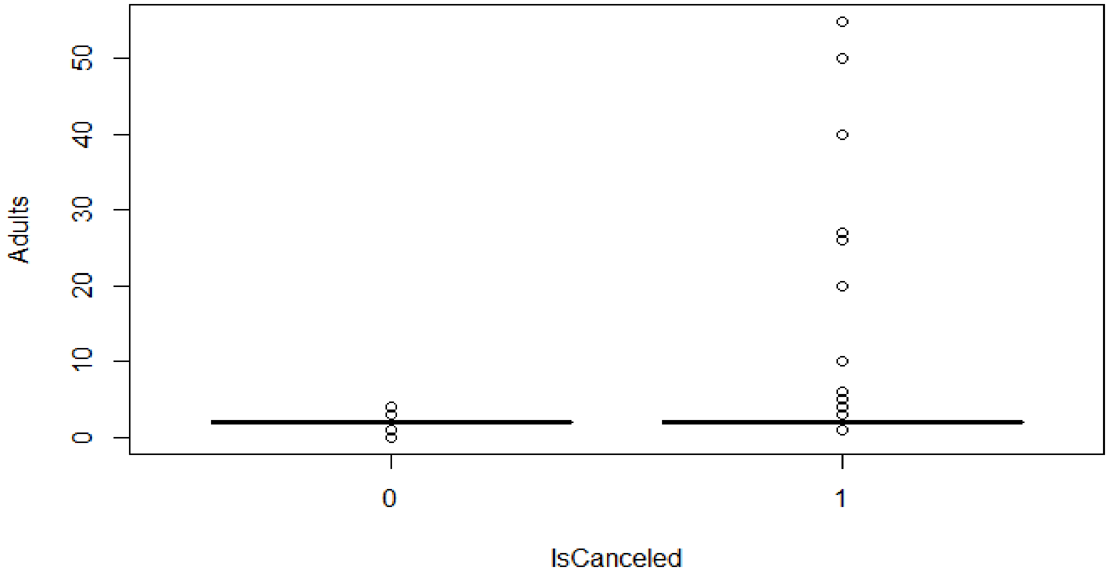


Fig. 8 Boxplot for adults and cancellations

This plot tells us that usually the reservations, which canceled, have less than ten adults and the majority of them had 2 adults.

Now we know that we should focus on people with non-refundable deposit types we plot the ten most important countries with the greatest number of reservations.

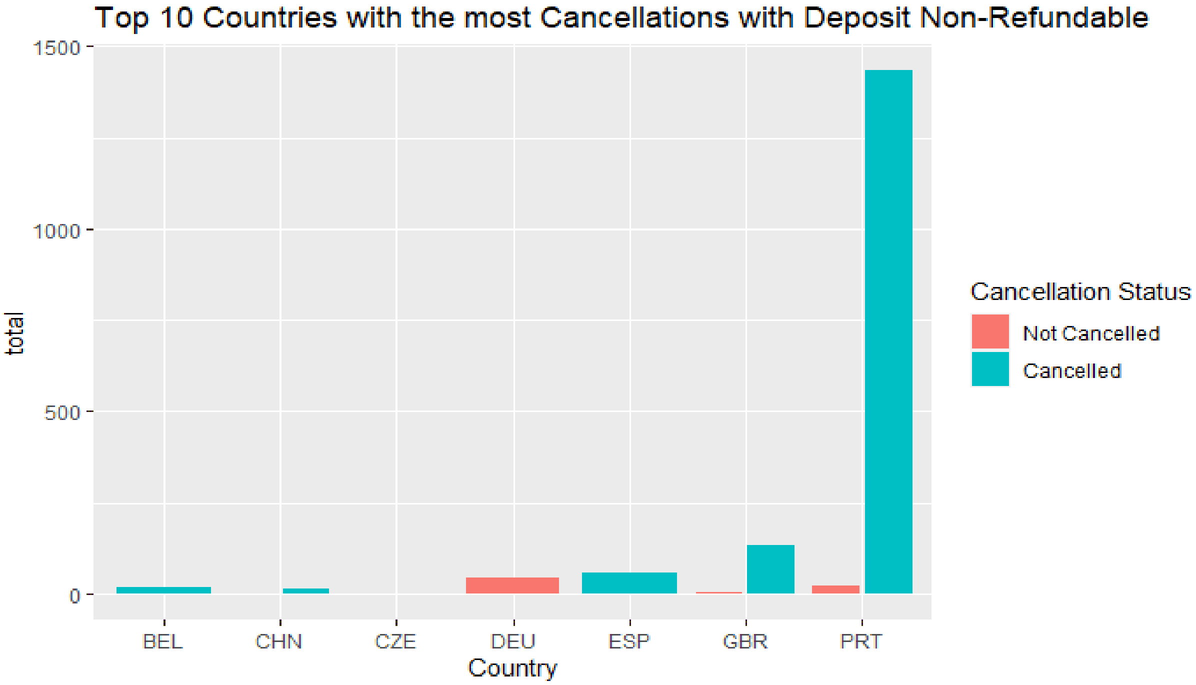


Fig. 9 Barchart for top 10 canceling countries

We can see that for people with a non-refundable deposit our target population is from Portugal.

From one of our regression trees, we can see some variables are more important to consider for example Lead Time, RequiredCarParkingSpaces, and BookingChanges.

For these variables, the model suggests that lead time should be separated into two groups. Those that have a lead time above 19 and those that have it less than that. Also, those that need parking spaces and those that don't. Moreover, those that have made booking changes and those that haven't.

Based on these variables we plot the number of cancellations by lead time and consider whether they needed parking spaces or not

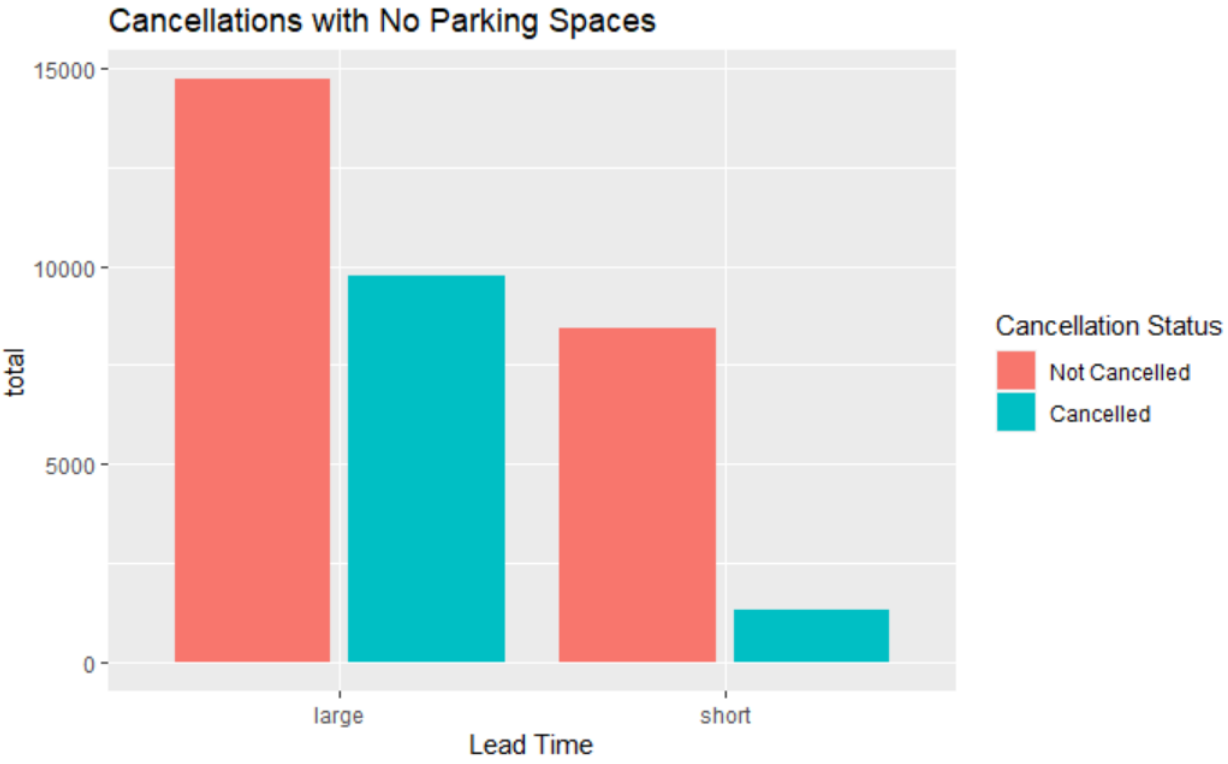


Fig. 10 Barchart for lead time and parking spaces

We can see that most people that don't require parking spaces had a large lead time, now we would like to answer the question of **why this is happening?** How we saw previously market segment is an important variable we should focus on we plot the number of cancellations of large lead time and consider whether they needed parking spaces by market segment.

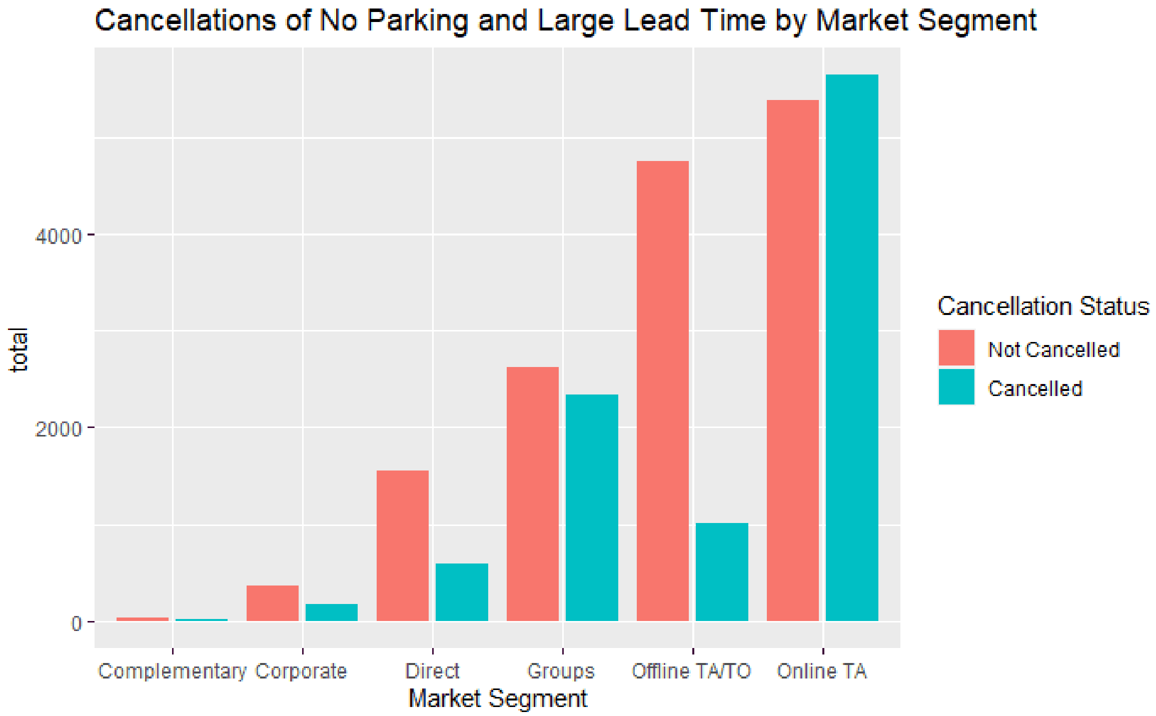


Fig. 11 Barchart based on the market segment

Now, if we focus on this group of people by the number of changes they have made

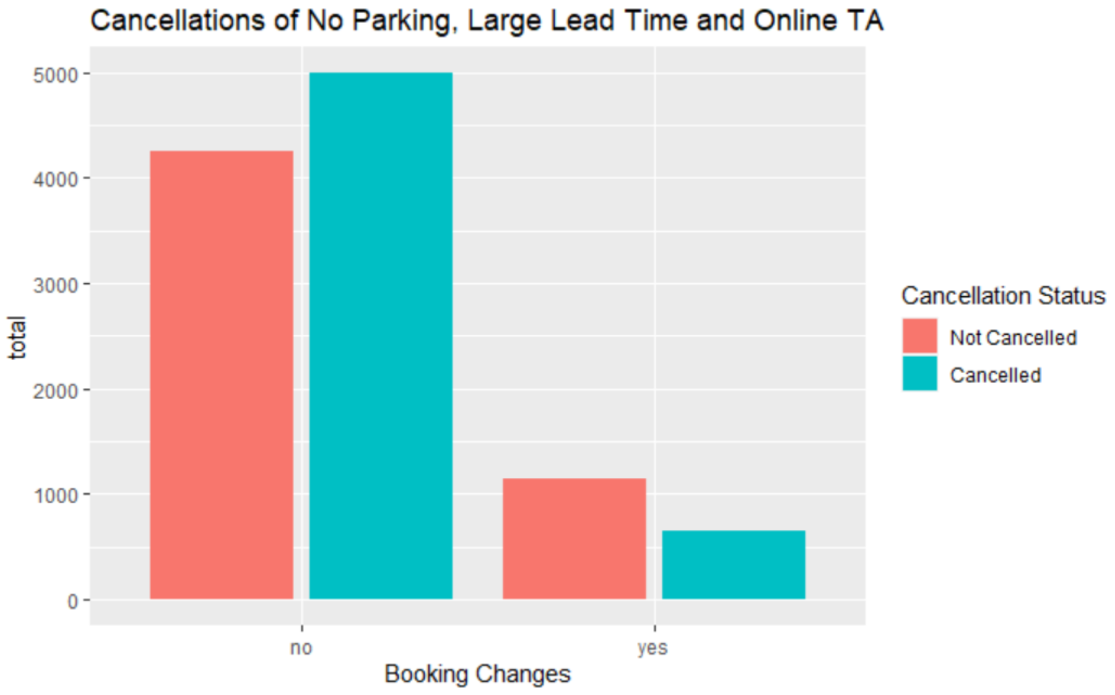


Fig. 12 Barchart based on booking changes

We can see from this plot, people with those characteristics didn't have booking changes at all. To solve the question of **why is this happening?** We decided to plot this by customer type because we saw previously that Transient people are a target market we should focus on.

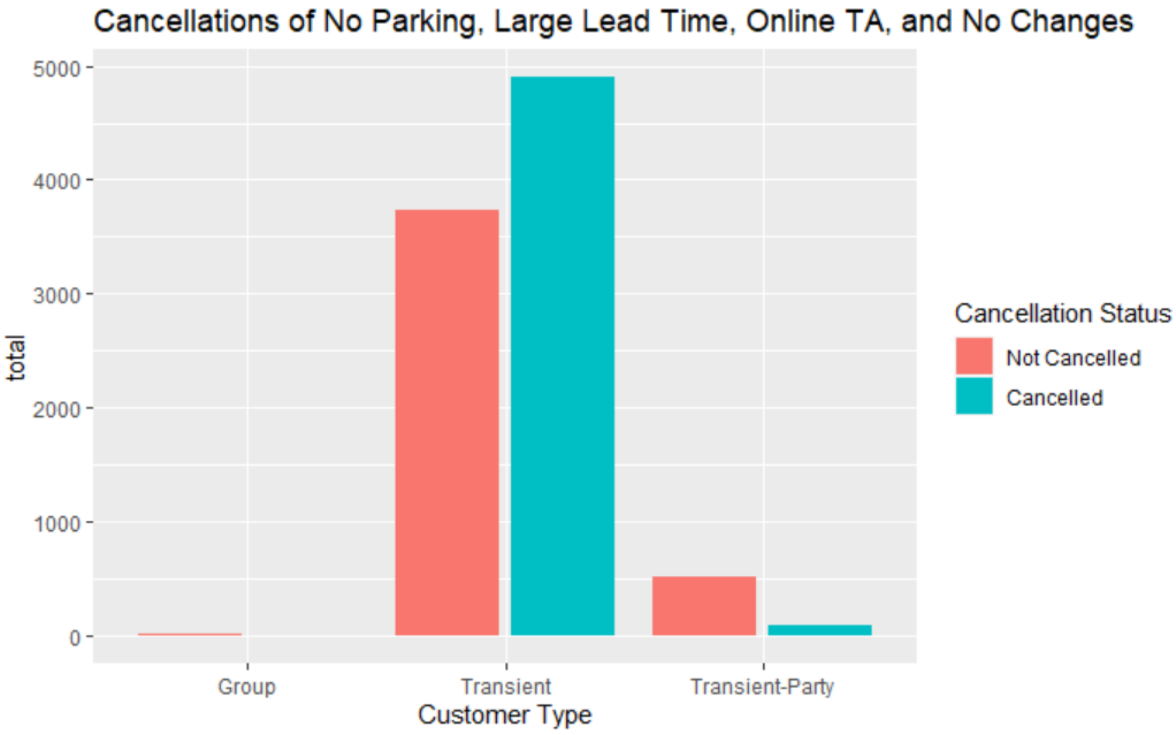
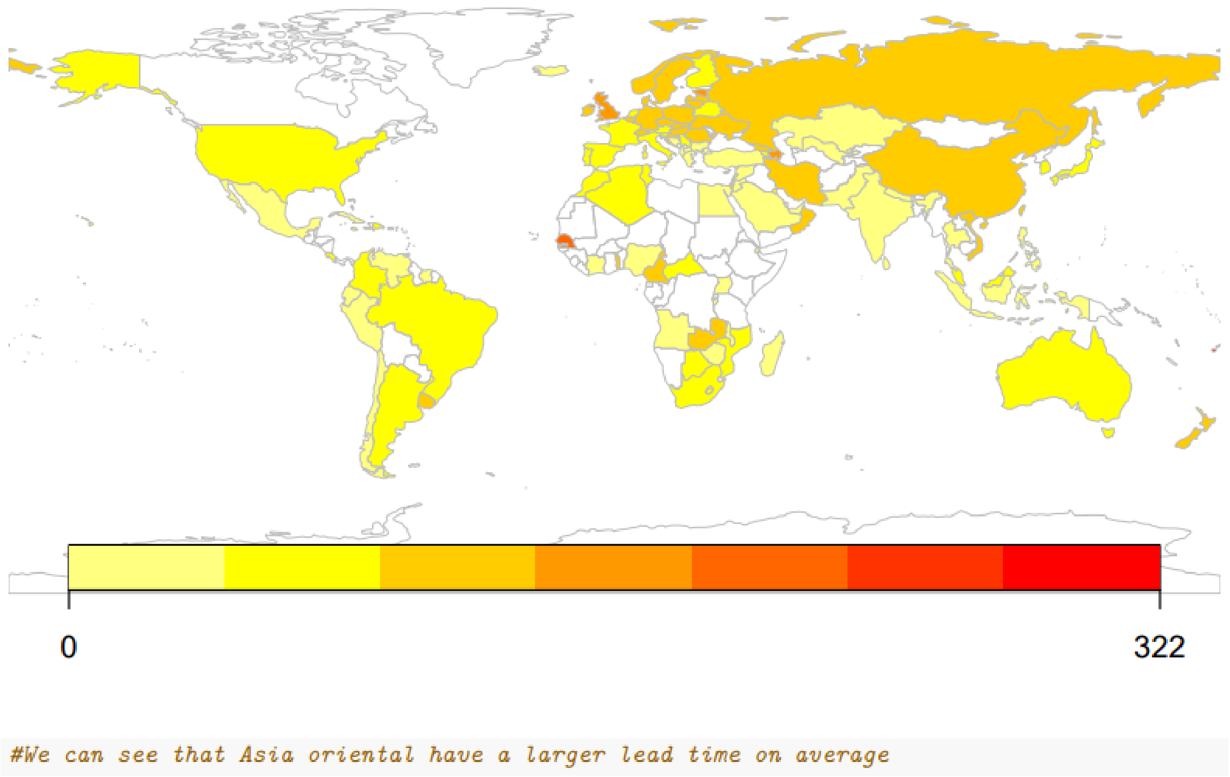


Fig. 13 Barchart based on customer type

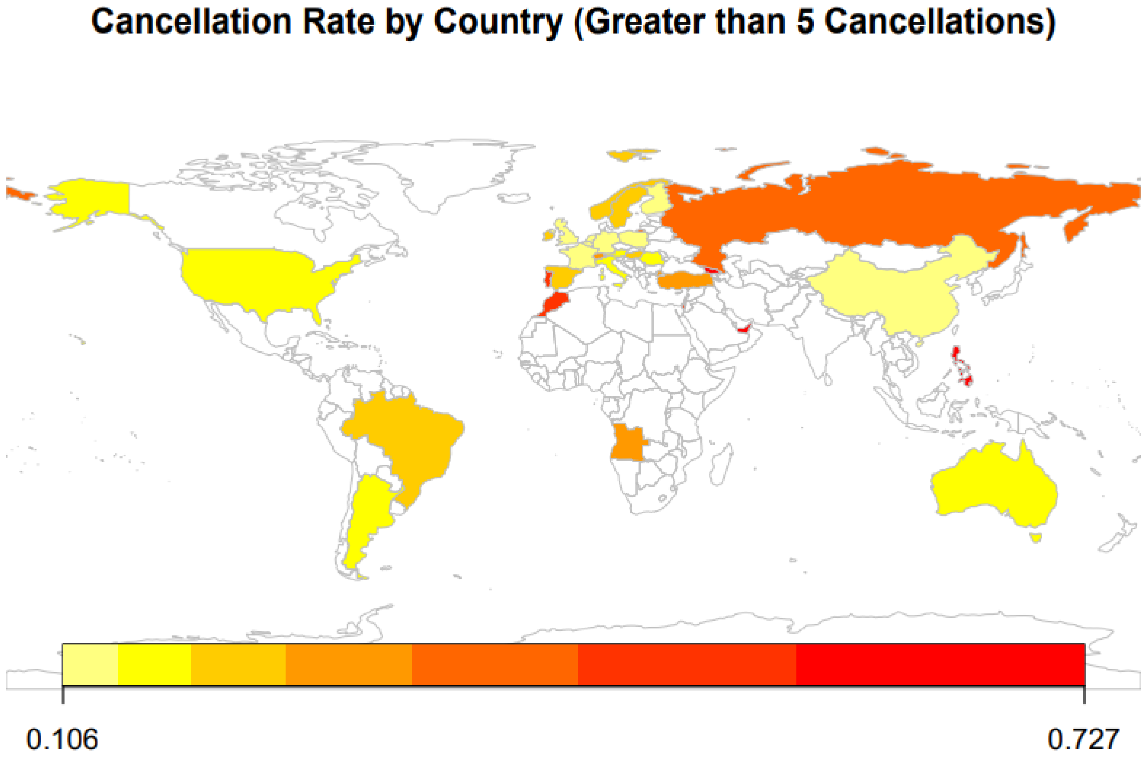
Now that we know that people with a large lead time, no parking spaces, from the Online TA market segment, that haven't made booking changes, and are considered transient have a large number of cancellations we will create association rules to know what is causing these people to cancel their reservations because they represent 44.23% of our cancellation population in the dataset.

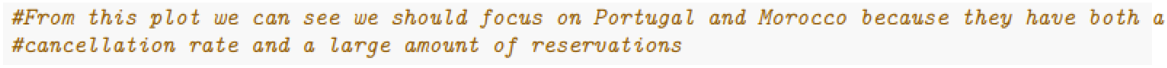
WORLD MAP ANALYSIS



```
#We can see that Asia oriental have a larger lead time on average
```

Fig 14: Lead time and the world



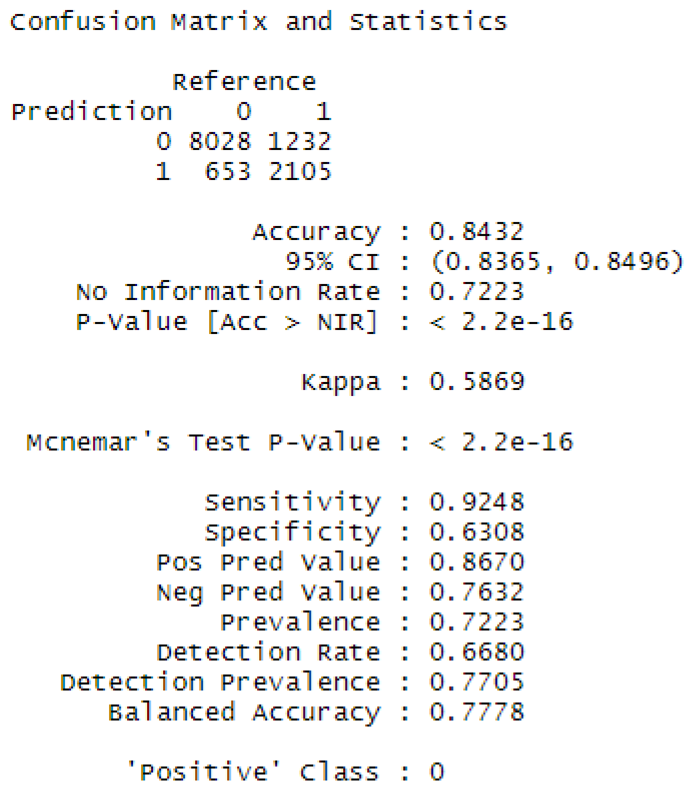


```
#From this plot we can see we should focus on Portugal and Morocco because they have both a high
#cancellation rate and a large amount of reservations
```

Fig 15 Countries and cancellations

LOGISTIC REGRESSION

Logistic Regression is one of the classification ML models and we pass the training data through the logistic regression model. family="binomial" as isCanceled is either '0' or '1'. It's a sophisticated statistical technique for modeling a binomial outcome using one or more explanatory variables. It estimates probabilities using a logistic function, which is the cumulative logistic distribution, to quantify the connection between the categorical dependent variable and one or more independent variables.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8028 1232
         1  653 2105
                                          
               Accuracy : 0.8432          
                 95% CI : (0.8365, 0.8496)
    No Information Rate : 0.7223          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.5869          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.9248          
            Specificity : 0.6308          
         Pos Pred Value : 0.8670          
         Neg Pred Value : 0.7632          
             Prevalence : 0.7223          
         Detection Rate : 0.6680          
   Detection Prevalence : 0.7705          
      Balanced Accuracy : 0.7778          
                                          
       'Positive' Class : 0               
```

Fig. 16 Confusion matrix statistics

RANDOM FOREST

A large number of decision trees are formed in the random forest approach. Every observation is input into the decision-making process. The final output is based on the most common conclusion for each observation. A new observation is fed into all the trees, with each categorization model obtaining a majority vote.

ntree - defines the number of trees to be generated. It is typical to test a range of values for this parameter (i.e. 100,200,300,400,500) and choose the one that minimizes the OOB estimate of error rate.

mtry - is the number of features used in the construction of each tree. These features are selected at random, which is where the “random” in “random forests” comes from. The default value for this parameter, when performing classification, is sqrt(number of features).

importance - enables the algorithm to calculate variable importance.

cutoff - Internally, random forest uses a cutoff of 0.5; i.e., if a particular unseen observation has a probability higher than 0.5, it will be classified as a positive class. In random forest, we have the option to customize the internal cutoff.

The Out-Of-Bag (OOB) data set is used to check the accuracy of the model, since the model wasn't created using this OOB data it will give us a good understanding of whether the model is effective or not.

Result: After Evaluating the probabilities and the class, the best random forest model gave an accuracy of 85.28% which is the best amongst all the models built.

REGRESSION TREE

A regression tree is a decision tree that is used for the task of regression which can be used to predict continuous-valued outputs instead of discrete outputs.

We created a regression tree considering the five more significant variables

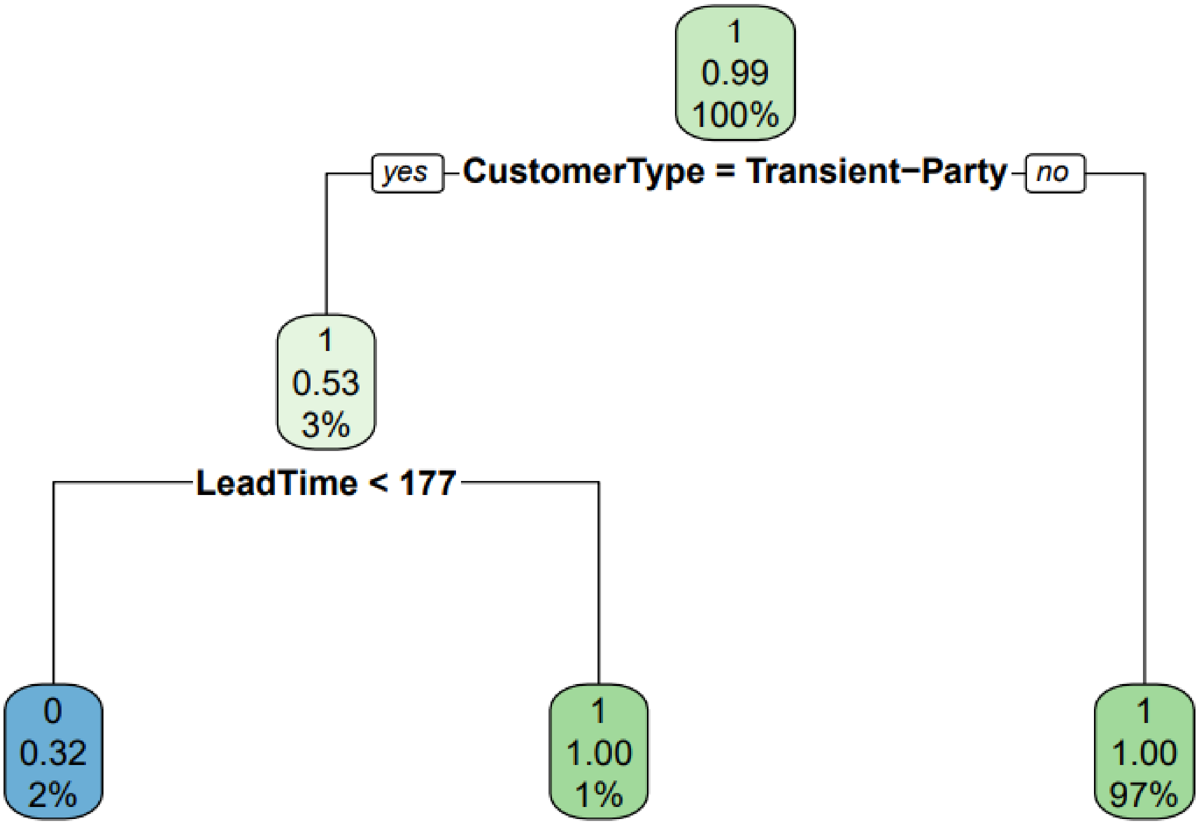


Fig. 17 Regression tree based on transient party and lead time

However, we see the model is not significant because the Mcnemar's Test P-Value is more than 0.05.

So, how the previous model wasn't significant. We started creating machine learning models to support our discovery that people from Portugal with a Non-Refundable deposit are people we should focus on.

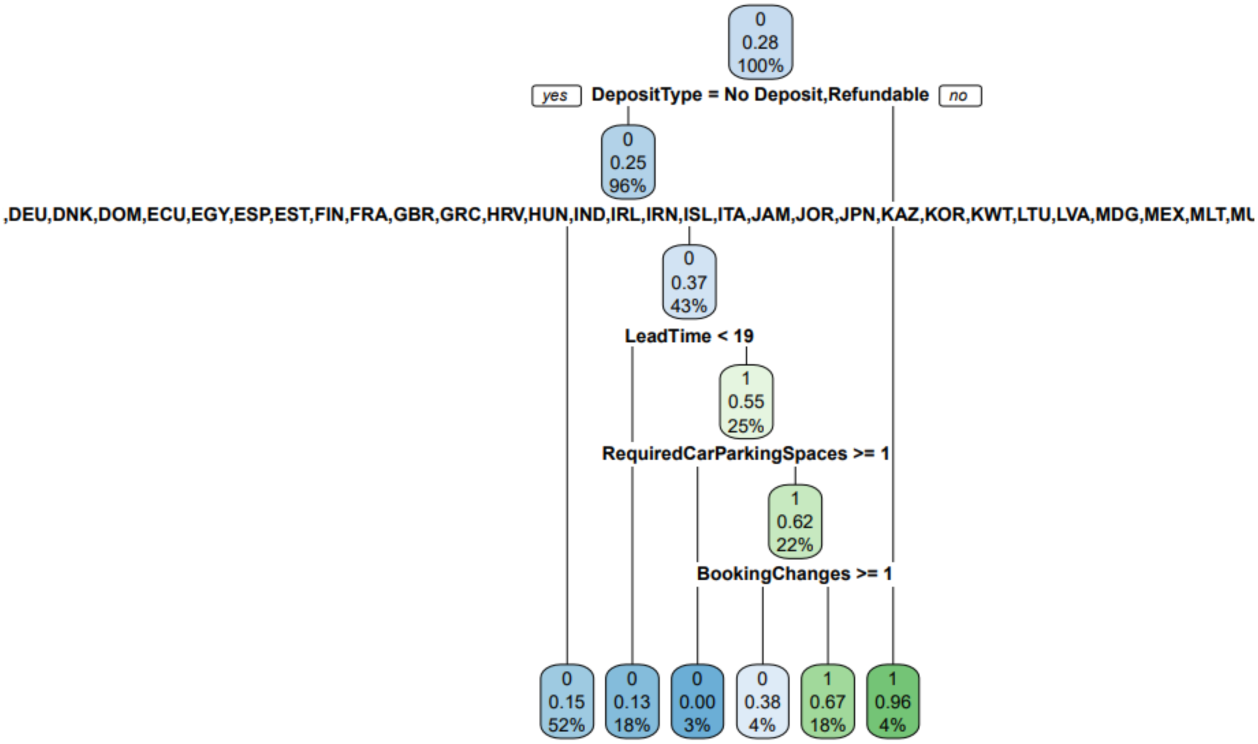
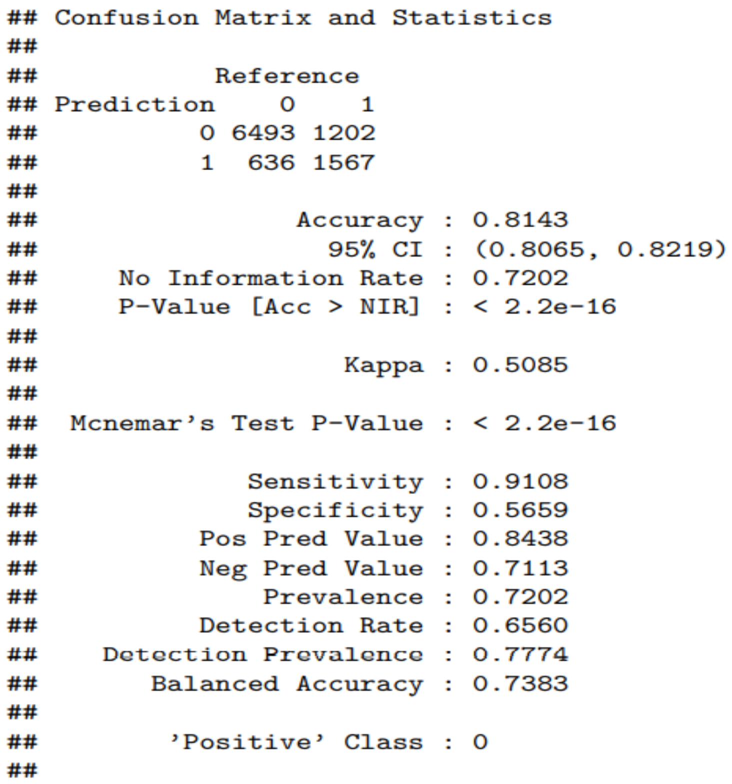

0
0.28
100%
yes
DepositType = No Deposit,Refundable
no
0
0.25
96%
,DEU,DNK,DOM,ECU,EGY,ESP,EST,FIN,FRA,GBR,GRC,HRV,HUN,IND,IRL,IRN,ISL,ITA,JAM,JOR,JPN,KAZ,KOR,KWT,LTU,LVA,MDG,MEX,MLT,ML
0
0.37
43%
LeadTime < 19
1
0.55
25%
RequiredCarParkingSpaces >= 1
1
0.62
22%
BookingChanges >= 1
0
0.15
52%
0
0.13
18%
0
0.00
3%
0
0.38
4%
1
0.67
18%
1
0.96
4%

Fig. 18 Regression Tree based on no deposit, refundable deposit with other aspects.



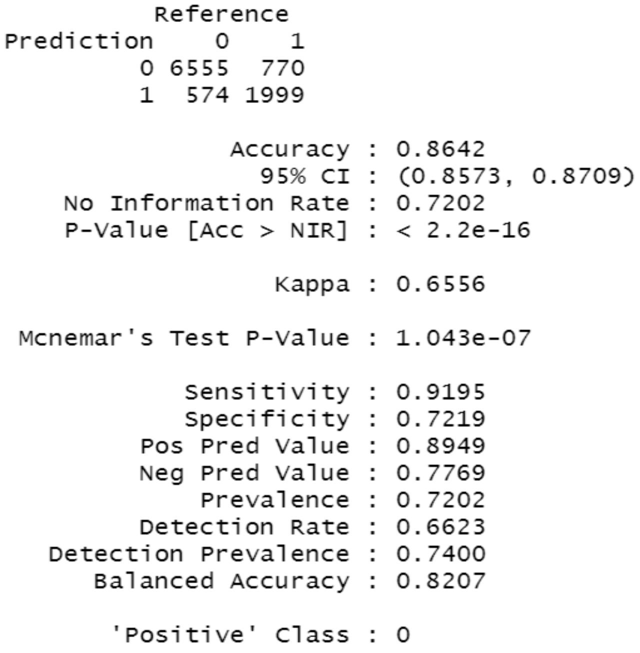
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6493 1202
##          1  636 1567
##
##                Accuracy : 0.8143
##                  95% CI : (0.8065, 0.8219)
##     No Information Rate : 0.7202
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5085
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9108
##             Specificity : 0.5659
##          Pos Pred Value : 0.8438
##          Neg Pred Value : 0.7113
##              Prevalence : 0.7202
##          Detection Rate : 0.6560
##    Detection Prevalence : 0.7774
##       Balanced Accuracy : 0.7383
##
##        'Positive' Class : 0
##
```

Fig. 19 Confusion matrix statistics

We infer that the model is significant because the Mcnemar's Test P-Value is less than 0.05. This model has an accuracy of 81.43%, which means the model predicted correctly the percentage of cases with the new data set.

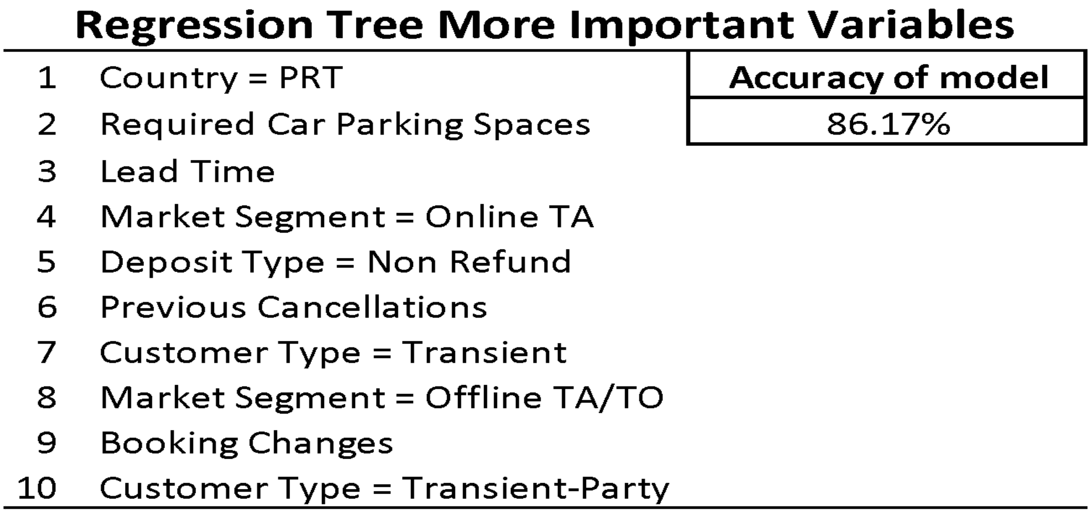
Moreover, if we compare sensitivity and specificity, we can conclude that the model is better at predicting when a person won't cancel (91.08% of the cases) than predicting when that person will cancel (56.59% of the cases).

We also did a more complex regression tree with cross-validation to support our findings. From the results, we can determine if people are going to cancel: if they come from Portugal, they have a non-refundable deposit, and the customer is transient.



```
          Reference
Prediction    0    1
         0 6555  770
         1  574 1999
                                          
               Accuracy : 0.8642          
                 95% CI : (0.8573, 0.8709)
    No Information Rate : 0.7202          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.6556          
                                          
 Mcnemar's Test P-Value : 1.043e-07       
                                          
            Sensitivity : 0.9195          
            Specificity : 0.7219          
         Pos Pred Value : 0.8949          
         Neg Pred Value : 0.7769          
             Prevalence : 0.7202          
         Detection Rate : 0.6623          
   Detection Prevalence : 0.7400          
      Balanced Accuracy : 0.8207          
                                          
       'Positive' Class : 0               
```

Fig. 20 Statistics



Regression Tree More Important Variables

1	Country = PRT
2	Required Car Parking Spaces
3	Lead Time
4	Market Segment = Online TA
5	Deposit Type = Non Refund
6	Previous Cancellations
7	Customer Type = Transient
8	Market Segment = Offline TA/TO
9	Booking Changes
10	Customer Type = Transient-Party

Accuracy of model
86.17%

Fig. 21 Important variables

For this model with cross-validations, we infer that the model is significant because the Mcnemar's Test P-Value is less than 0.05. This model has an accuracy of 86.42%, which means the model predicted correctly the percentage of cases with the new data set.

Therefore, this model supports our idea that people from Portugal and with a non-refundable deposit have a high probability to cancel.

ASSOCIATION RULE MINING

Association rule mining is a procedure that aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.

From the association rules
We should focus on people that come from Portugal, have a non-refundable deposit, and are considered transients. This is because they represent 12.67% of all the cancellations in the dataset.

We can observe that the bookings are the highest when done through an online Travel Agent. It also guarantees that through online travel agents, there is a greater probability of no cancellations.

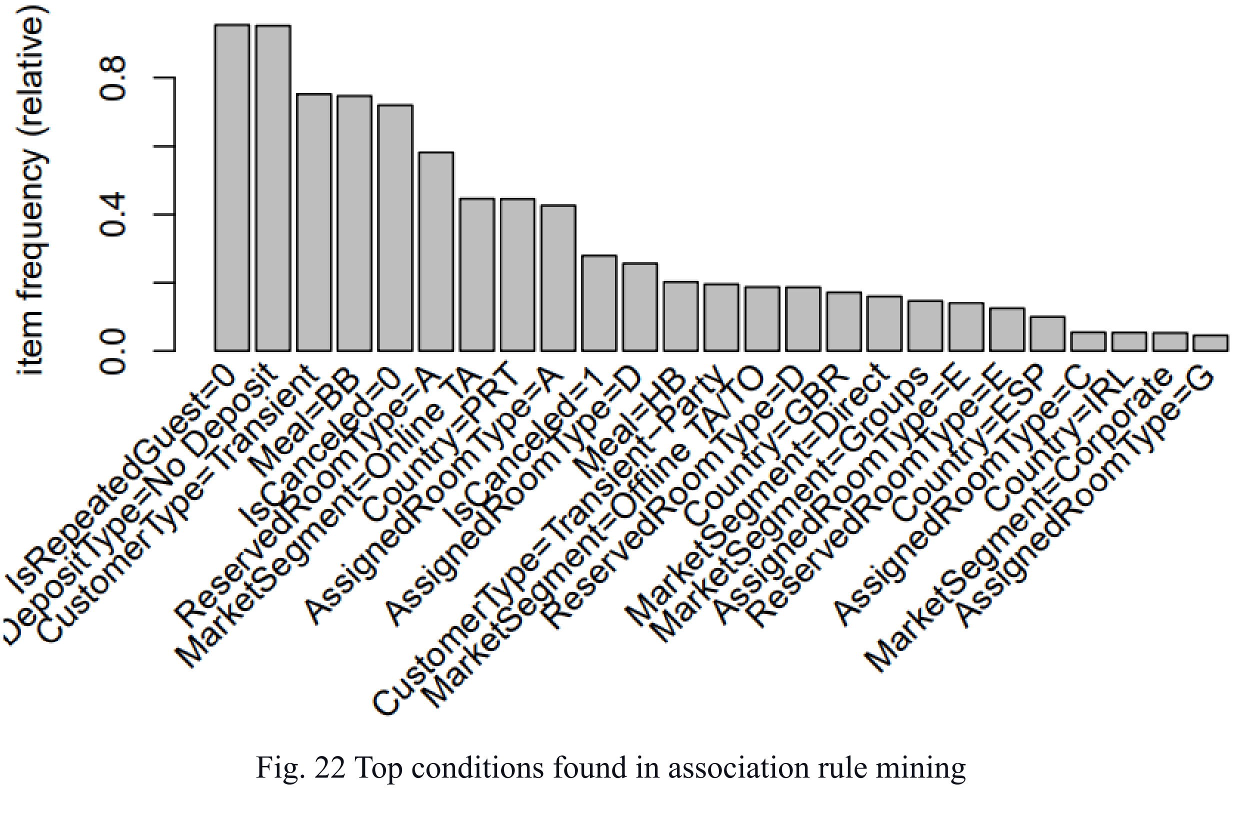
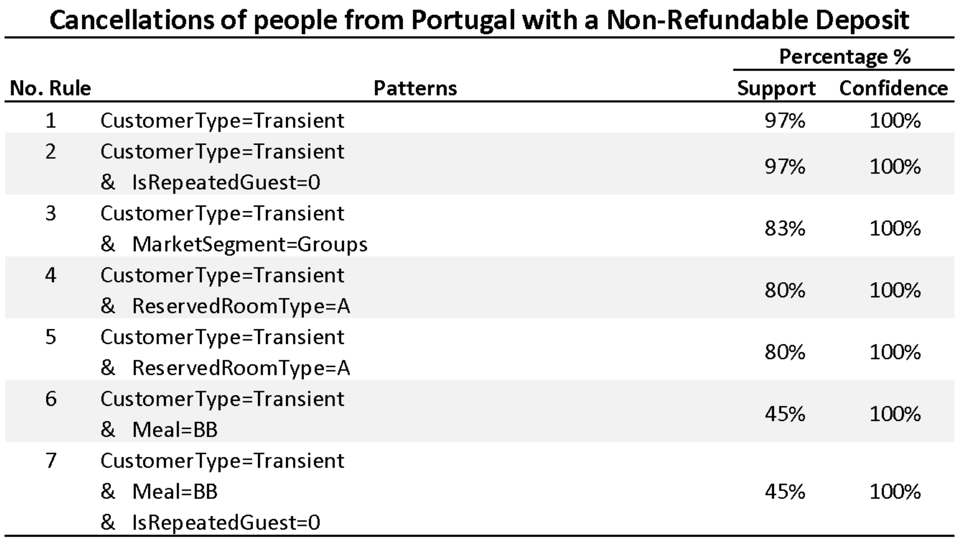


Fig. 22 Top conditions found in association rule mining

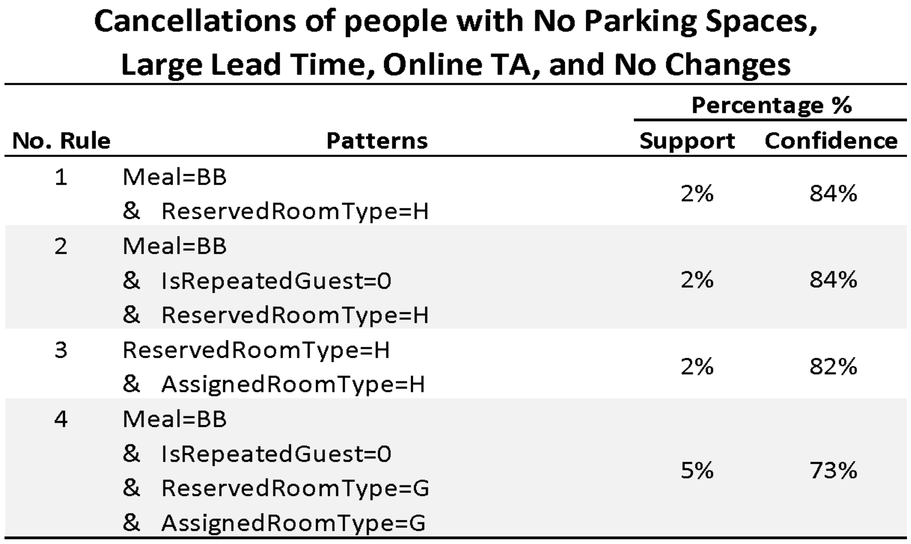


Cancellations of people from Portugal with a Non-Refundable Deposit

No. Rule	Patterns	Percentage %	
		Support	Confidence
1	CustomerType=Transient	97%	100%
2	CustomerType=Transient & IsRepeatedGuest=0	97%	100%
3	CustomerType=Transient & MarketSegment=Groups	83%	100%
4	CustomerType=Transient & ReservedRoomType=A	80%	100%
5	CustomerType=Transient & ReservedRoomType=A	80%	100%
6	CustomerType=Transient & Meal=BB	45%	100%
7	CustomerType=Transient & Meal=BB & IsRepeatedGuest=0	45%	100%

Fig. 23 Association rule mining stats

From our association rules model for people that come from Portugal and have a non-refundable deposit type, we can see that if the person is also transient, there is 100% confidence that he/she is going to cancel the reservations with a support of 97%. Moreover, most of the rules have room type A, so this can suggest there is something wrong with that bedroom. Actually, in general, we saw in a previous graphic that people that stay in room A cancel more times than other people staying in other rooms.



Cancellations of people with No Parking Spaces, Large Lead Time, Online TA, and No Changes

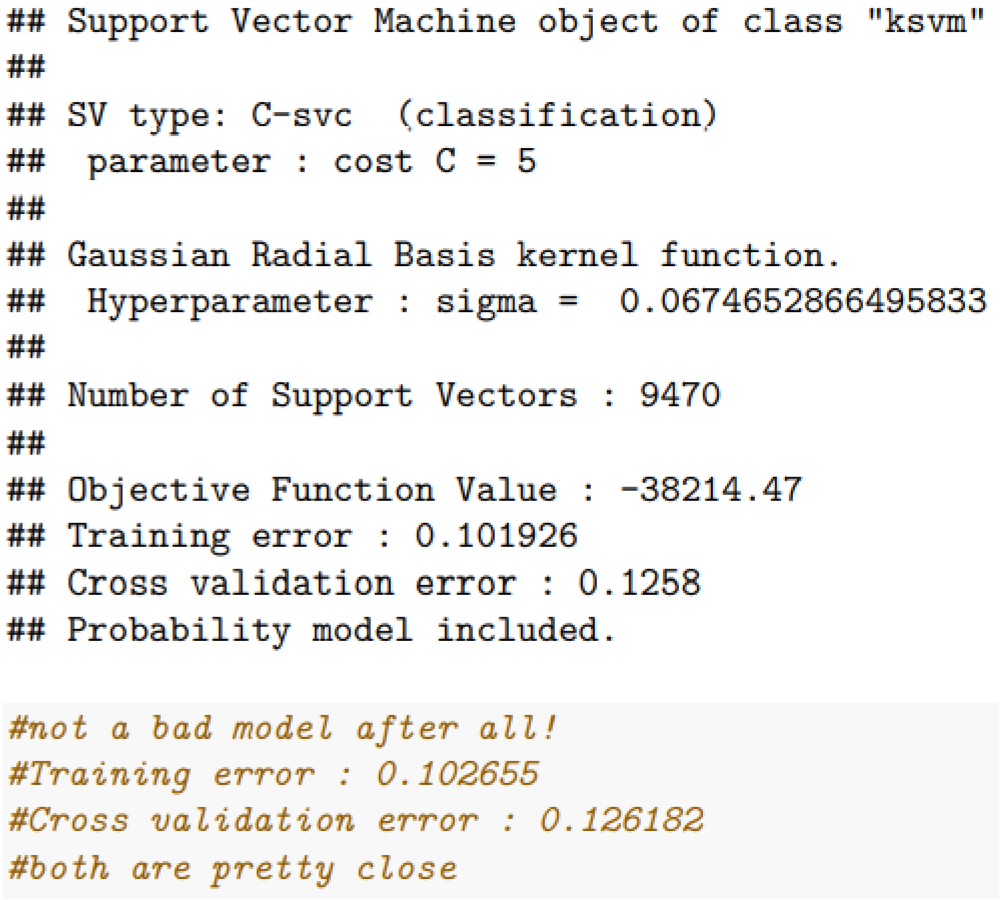
No. Rule	Patterns	Percentage %	
		Support	Confidence
1	Meal=BB & ReservedRoomType=H	2%	84%
2	Meal=BB & IsRepeatedGuest=0 & ReservedRoomType=H	2%	84%
3	ReservedRoomType=H & AssignedRoomType=H	2%	82%
4	Meal=BB & IsRepeatedGuest=0 & ReservedRoomType=G & AssignedRoomType=G	5%	73%

Fig. 24 Association rule mining stats

The results of association rules for people with a large lead time, no parking spaces, from the Online TA market segment, that haven't made booking changes, and are considered transient suggest we should focus on Meal plan BB and Room H because there is an 84% they will cancel their reservations if they meet this features.

SUPPORT VECTOR MACHINES

Support vector machines (SVMs) are supervised learning models that examine data for classification and regression analysis in machine learning.



```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 5
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.0674652866495833
##
## Number of Support Vectors : 9470
##
## Objective Function Value : -38214.47
## Training error : 0.101926
## Cross validation error : 0.1258
## Probability model included.
```

```
#not a bad model after all!
#Training error : 0.102655
#Cross validation error : 0.126182
#both are pretty close
```

Fig. 25 SVM statistics

Insights:

- Maximum cancellations came from Portugal
- Direct market and Offline TA has lesser cancellations
- Online TAs had the most cancellations
- Bookings with children lead to lesser cancellations
- The higher lead time tends to have lesser cancellations
- No deposits promote cancellations
- When a parking spot is reserved, the probability of cancellation decreases significantly
- Room type ‘A’ had the highest number of cancellations
- Maximum cancellations of room A bookings are done online
- Repeated guests tend to cancel less frequently
- Portugal and Morocco have large reservations but also high cancellation rates

RECOMMENDATIONS

- Change deposit type for bookings from Portugal
  A majority of our bookings are based out of Portugal. Those with a non-refundable deposit and transient customer type constitute 12.67% of the canceled population.

- Upgrade rooms or provide options for room switching.
  Since Room Type A is highly linked with cancellations, people who have booked room A could be provided an option to choose another room type on a last-minute basis.

- Improve graphics or data available for room type A online
  Since the majority of the cancellations of room type A are through online TA, maybe the customers had expected better rooms and were disappointed on seeing the actual room. Updating the room pictures provided to the online platforms, and keeping the rooms up to the mark may help.

- Special offers for group bookings with deposits
  Encourage customers of group type bookings to opt for deposit type bookings with special discount offers, provide gift vouchers or introduce a points-based system to avail discounts for future travel plans.

- Promote parking spaces and increase their availability:
  Since bookings with parking requirements have proved to be successful, we suggest advertising the availability of parking spaces effectively.

- Celebrate repeated customers:
  Repeated customers have proved to be loyal customers and rewarding them with special discounts or cashback or even a loyalty program seems like a good strategy.

- Breakfast plans:
  Since Room Type H and Meal Plan HB is highly linked with cancellations, people of the above subtype who have booked Room Type H or Meal Plan BB could be provided an option to choose another Room Type or Meal Plan on a last-minute basis.
  Generally, people tend to book BB plans for travel and holiday packages. We see a scope of improvement here.