**Hotel Booking Dataset Analysis Report**

Group 5 Member: Achal Suresh, Aditi Baijal, Manyue Chen, Baojing Gou, Yaosong Yu

**Business Understanding**

The profitability of a hotel is related to many factors, but the most important one is occupancy, and since hotels have a limited number of rooms available at any time proper management of the number of rooms is needed for maximizing revenue. And with more and more hotels accepting bookings in advance last-minute cancellations can result in lost revenue for the hotels. WIth cancellations reaching figures as high as 25% (Antonio, Almeida & Nunes,2017 ) the cancellations are definitely resulting in significant losses to hotels.

There are a number of reasons that can result in cancellations such as business meetings changes, vacations rescheduling, illness, bad weather conditions, customer behavior such as looking for better deals, etc. also result in cancellations. And as a solution to these hotels have started different mechanisms such as cancellation charges, overbooking the hotel, etc., however, this has not deterred cancellations. And moreover, these measures can also affect the relationship with customers who might feel deterred to make advanced booking due to fear of cancellation charges.

Other measures taken by hotels are providing discounts and other services on booking being made, providing personal interaction with customers are other measures that have been taken by hotels to prevent customers from making cancellations.

To counter such scenarios utilizing data science models to better predict the chances of booking cancellations can be helpful to both customers to better protect themselves from the last minutes and hence reduce uncertainty for both parties. At the same time, developing a classification prediction model enables hotels to act upon those specific bookings to try to avoid their cancellation and help hotels to target only customers who will show less likelihood of canceling rather than all guests. Such targeted marketing can be highly beneficial to the organization.

With the huge amount of data being generated about the booking patterns as well as customer behavior data science now provides an ideal scenario to better understand the intricate patterns in bookings and cancellations. For our study, we have taken the data from four hotels/resorts in Portugal between the years 2015-2017 for analysis.

Our group would like to use this data to analyze what factors affect occupancy most, to know what types of customers are more likely to cancel their reservations. The model can also help hotels to keep more rooms available for rooms being vacant because of future cancellation. We expect to reach a conclusion, hoping to help the hotel to determine in advance the different needs of the canceled guests and the guests who have already checked in. Reach the right audience with time-based marketing by targeted advertising to guests with high occupancy rates. The team aims to utilize different modeling techniques such as Random Forest, Logistic Regression, Decision Tree and Naïve Bayes to determine which model can predict better and hence provide a more suitable forecasting technique

**Data Understanding & visualization**

Dataset

The data was retrieved from Kaggle <https://www.kaggle.com/jessemostipak/hotel-booking-demand> and illustrated the records from 4 hotels in Portugal from 2015-2017

The data set contains roughly 119390 instances which is a good amount of volume that can be split into training and test data for model generation. The description for each of the variables and its description in the appendix Exhibit A.

Target Variable

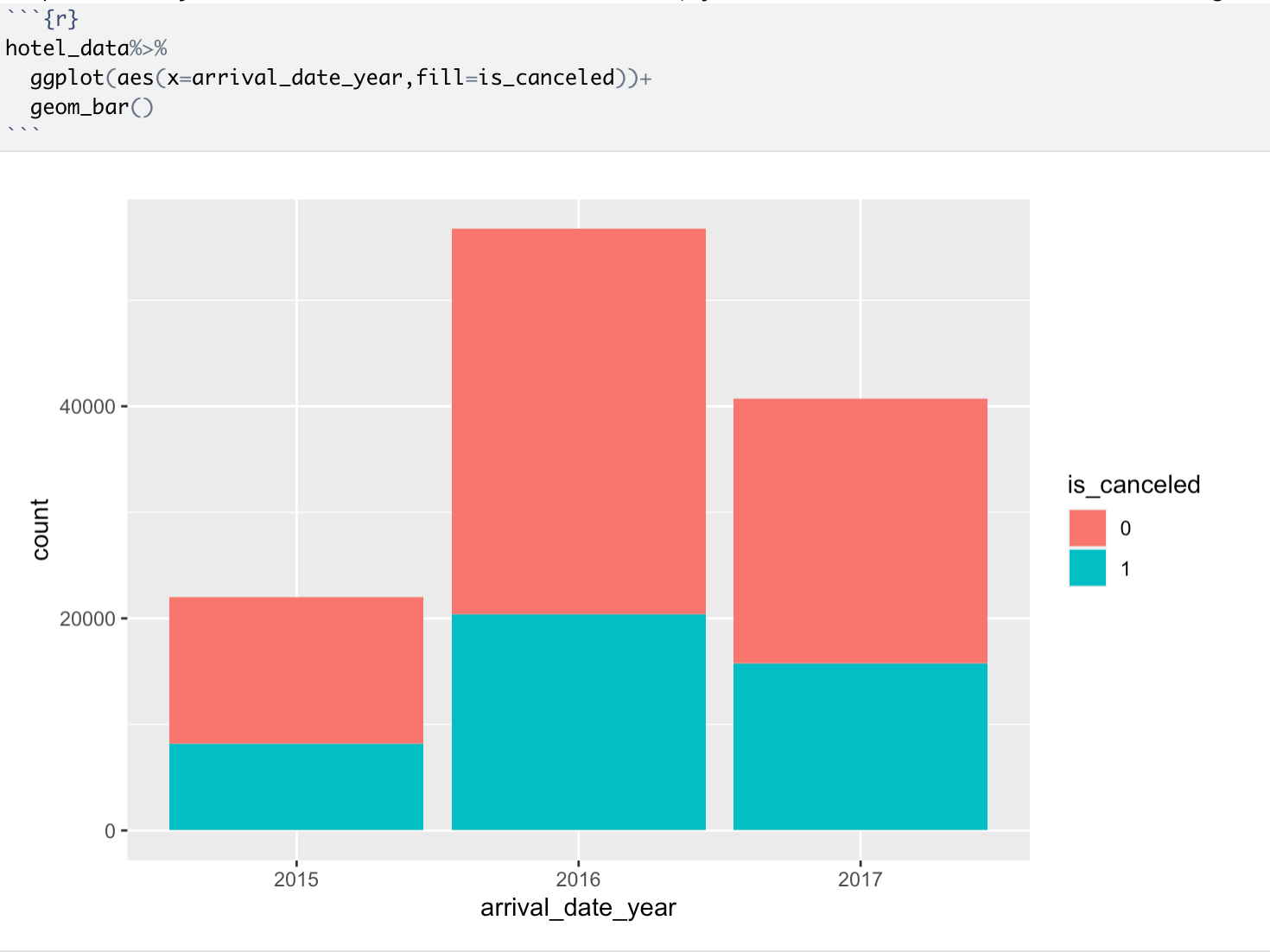
The is\_cancelled would be the target variable. We intend to predict whether a hotel booking will likely be canceled or not

Through our initial analysis, we saw that there are a large number of null values in the columns Agent and Company had a large number of columns. Also, columns such as reservation status have values similar to our target variable ‘is\_canceled’. We also had to convert a number of columns such as hotels, meals, etc to factors for our analysis.

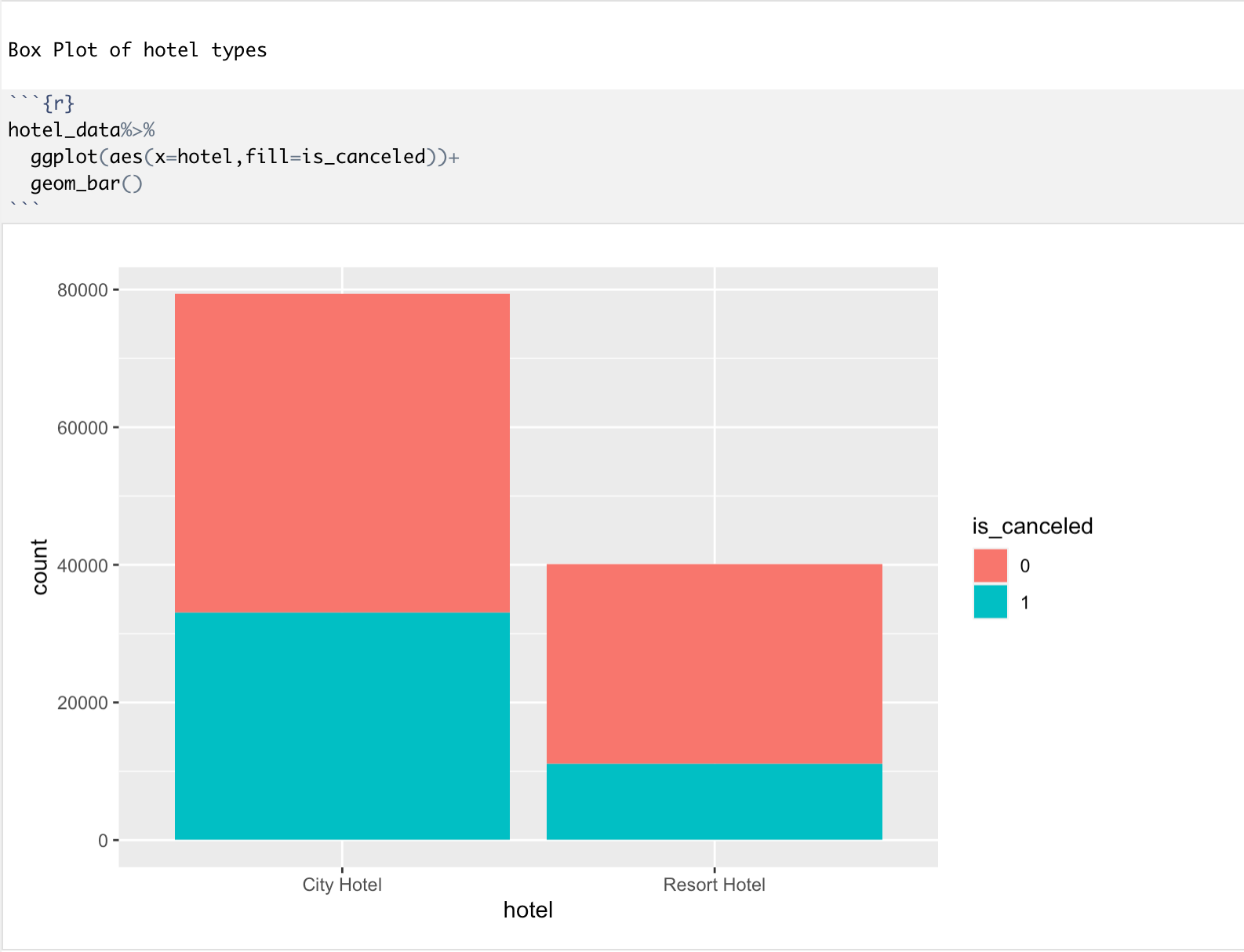
We have also created two new columns stay nights total = weekend nights + weekday nights and stay total cost = stay nights total \* cost per night (adr).

Our group select **Hotel, Stays\_in\_weekend\_nights, Stays\_in\_week\_nights , Adults,** **Children, Babies , Market\_segment, distribution\_channel,distribution\_channel , Is\_repeated\_guest , Previous\_cancellations , Booking\_changes , Deposit\_type ,Days\_in\_waiting\_list are the main** *features that are gonna be used to create the model.*

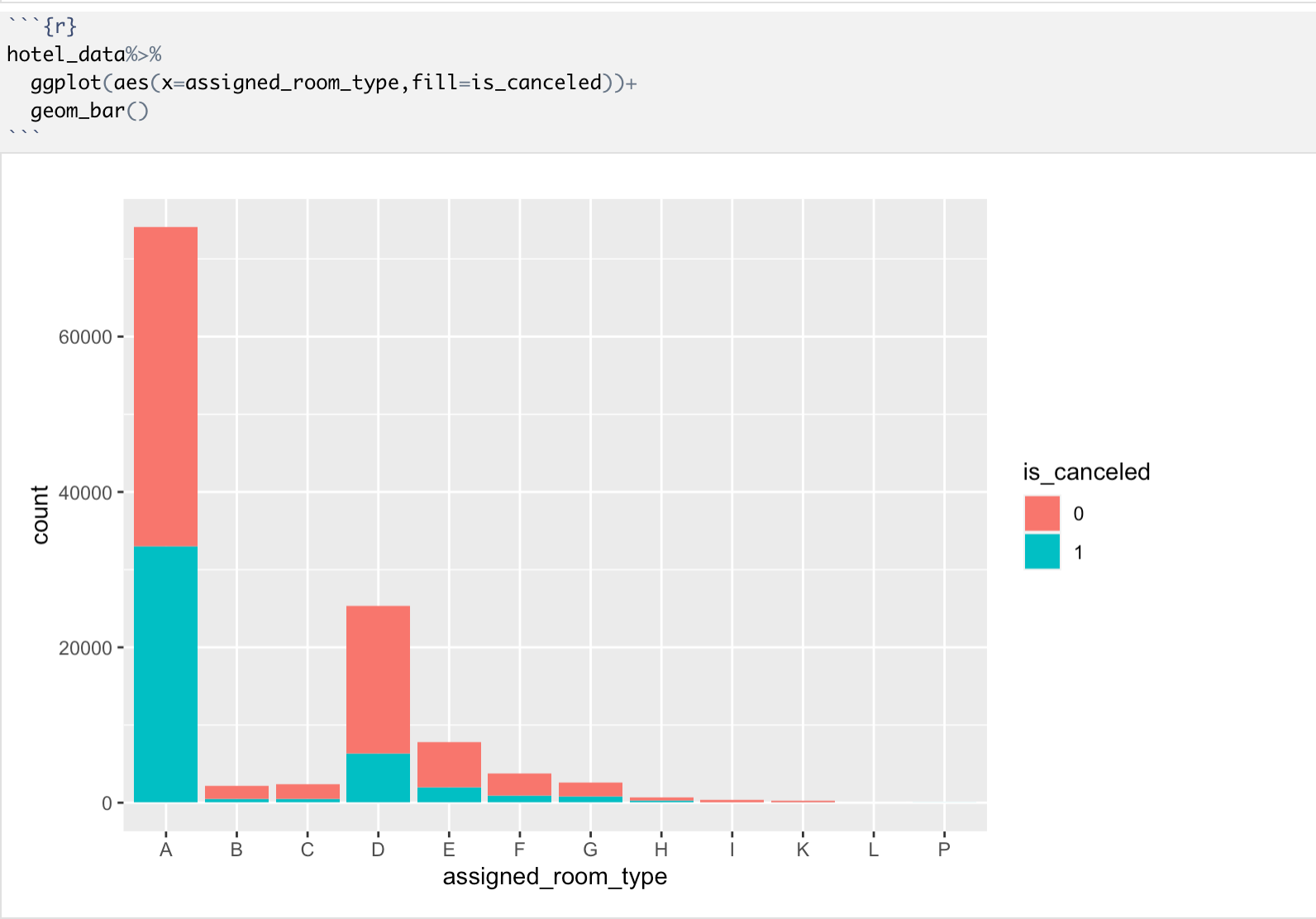
We did data visualization for exploring several interesting patterns in this dataset:



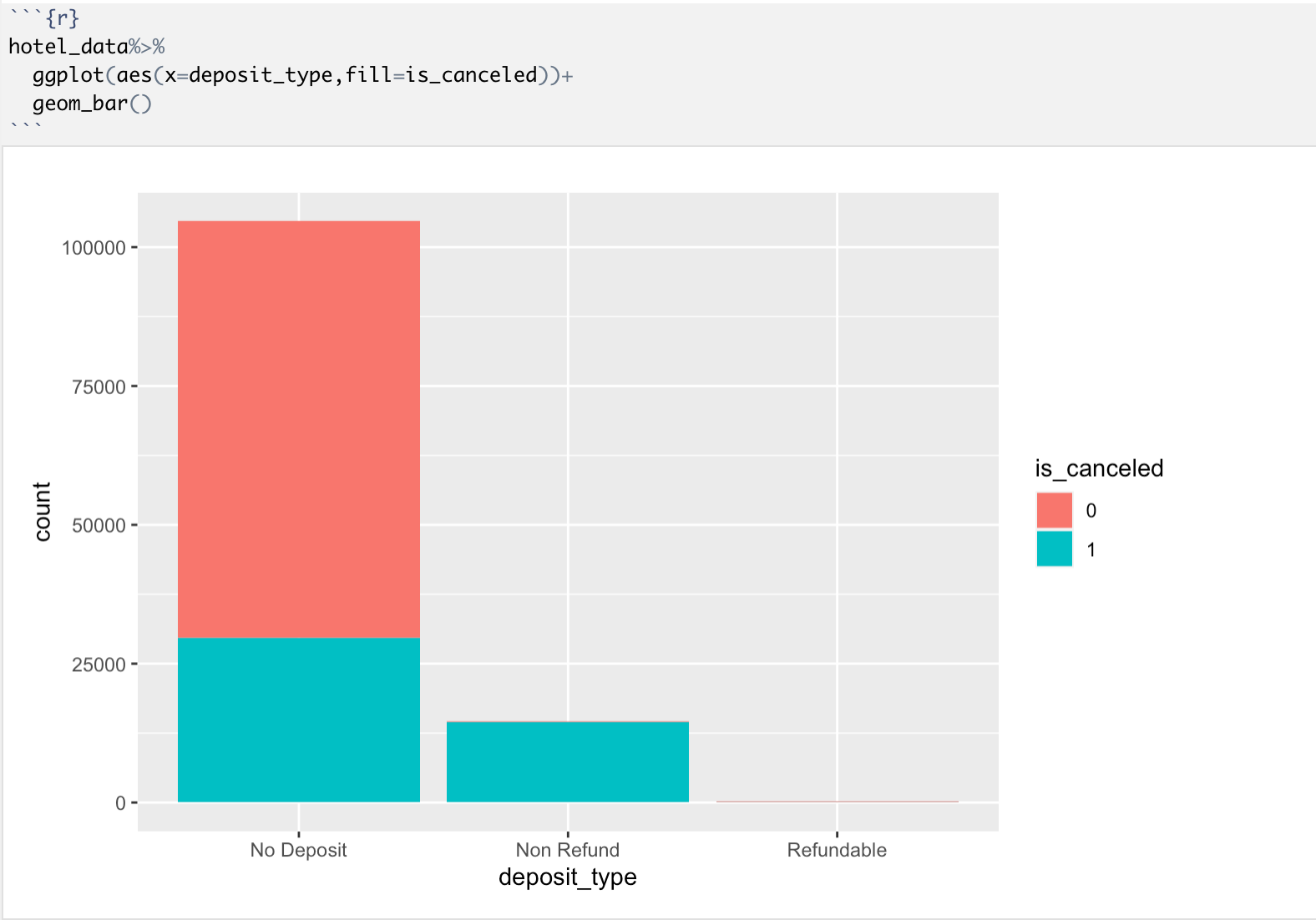
We can see that the highest number of booking and cancellations happened in 2016 and the lowest number of booking and cancellations happened in 2015.



We can observe from this graph that the booking number of city hotels is almost 2 times more than that of resort hotels(8000 compares with 4000), and the cancelation number of city hotels is almost 3 times more than that of resort hotels(3000 compares with 1000).



This graph visually shows which type of room types might receive the highest cancellation. Most cancellations come from type A room and type D. Nearly 45% of bookings from type A room would like to be canceled(7400 compares with 3400).



This graph is unexpected. People often feel that a non-refundable deposit will help prevent a cancellation. However, in this dataset, almost all non-refundable deposits were cancelled, although the total number of cancellations from no-deposit was the highest.

**Modeling**

Logistic Regression

From the data understanding and visualization phase, we saw that some of the columns are showing the relationship with the cancellation variable. So for our modeling, we decided to utilize the following variables:

hotel, is\_canceled, lead\_time, adults, children, babies, meal, market\_segment, distribution\_channel, is\_repeated\_guest, previous\_cancellations, previous\_bookings\_not\_canceled, reserved\_room\_type, deposit\_type, days\_in\_waiting\_list, customer\_type, adr, required\_car\_parking\_spaces, stay\_nights\_total, stay\_cost\_total

For our first model, we decided to utilize logistic regression to model the data. Logistic regression presents a simpler classification algorithm that can also help understand the dependence of different variables.

We split the data into training and testing set in the ratio of 70:30 for evaluating the model.

Some of the observations from the model based on the output of the logistic regression as shown in Exhibit B

* With an increase in the lead time to the arrival date of the guests, the odds of cancellation is also showing a significant increase.
* Guests having babies are showing lesser odds of cancellations while guests that have children have shown a very high increase in the number of cancellations
* Repeated guests are showing a lower tendency to make cancellations while individuals who have previously made cancellations show the likelihood to make cancellations again.
* Individuals who make reservations with refundable deposit type show the likelihood of making cancellations.
* An increase in the number of days booking and total costs will also lead to guests making cancellations.

Classification - Decision Tree

Following steps are followed in order to obtain best tree model:

1. build a model using a training set (80% of the overall data). We now have 20% of the dataset that is not used to build this model but rather test it.
2. The model will be evaluated on test data, it can be done by providing a dataset name with a model to `predict()` function.
3. Using `plotcp()`, check for cross-validation error rate change based on complexity of the model increase.`xerror` is cross-validated relative error rates, and `xstd` is its standard deviation.
4. CP with lowest xerror value is used to obtain the best model

Model Observations:

1. Using validation the CP value with lowest error is 0, which is used for modeling
2. Variables that play key in tree model are : deposit\_type, lead\_time, customer\_type, previous\_cancellations and hotel\_type.
3. The probability of booking cancellation increases given if it was booked under no refund deposit type with city hotel type and customer type other than transient. Cancellation probability is also high with deposit type either no deposit or non refundable and lead time greater than 12 and previous cancellations present.

Exhibit C showcases the output of Decision Tree.

Random Forest

We use Random Forest to classify the probability of cancellation for each hotel and do votes from each tree.

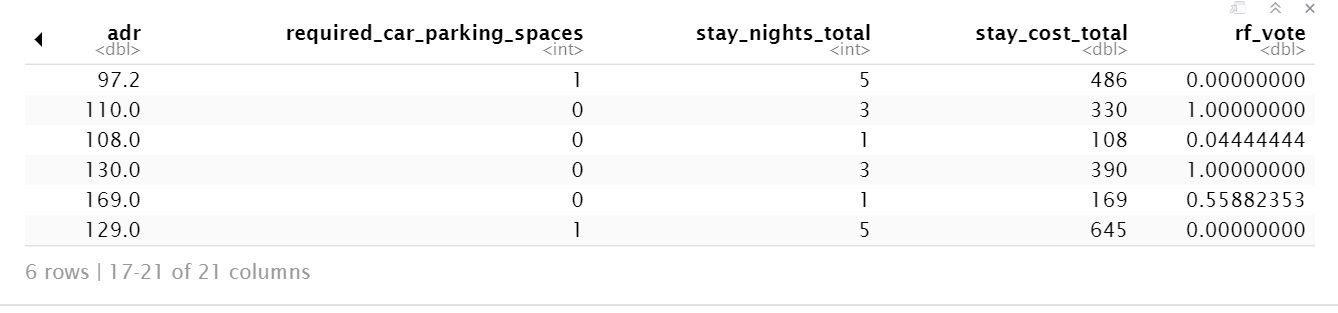
Since the Rstudio cannot run a random forest model with more than 53 categories, we look at the levels of variables and find out that variables “country”, “agent”, “company” each have more than 100 levels, so we remove these three variables.

The final data we use is: is\_canceled (“1”=canceled; “0”=No cancled), hotel, lead\_time, adults, children, babies, meal, market\_segment, distribution\_channel, is\_repeated\_guest, previous\_cancellations, previous\_bookings\_not\_canceled, reserved\_room\_type, deposit\_type, days\_in\_waiting\_list, customer\_type, adr, required\_car\_parking\_spaces, stay\_nights\_total, stay\_cost\_total.

From the output of the random forest model:

From the Mean Decrease Accuracy plot (Exhibit D), lead\_time, with a value over 200, has the most influence on the accuracy of the model, deposit\_type is the second most important factor to influence the accuracy of the model. Babies variable has the least importance to the accuracy of the model.

The Mean Decrease Gini plot (Exhibit D) shows that deposit\_type and lead\_time have the most importance with a value of over 6000 in the model, the stay\_cost\_total and adr\_ are the second most important with a value of around 6000. These two factors have the greatest effect on the random forest model, while the Babies and is\_repeated\_guest are the least important factors with a value of 0.



We can have the possibility of cancellation or no cancellation for each hotel. For example, the probability of the 1st hotel to be canceled by a customer is 0, and the probability of the 2nd hotel to be canceled is 1.0, and so on and so forth.

Random forest can have a very good predictive performance, but it almost shows all the situations that lose rules to explain.

Naive Bayes

We, therefore, considered Naive Bayes model because NB assumes independence among variables. Observing that the database covers many different variables, we presumed that independence can be assumed. Before running the NB model, we predicted that this model will be the most accurate one.

**Evaluation**

Logistic Regression

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| Actual/  Predicted | 0 | 1 |
| 0 | 21787 | 6989 |
| 1 | 858 | 6182 |

Accuracy of the model defined as the total number of Accurate Predictions/ Total number of predictions

78.09% which indicates that the logistic regression is able to create a pretty accurate model for analysis. However, there is a large number of False Negatives being shown in the model which can affect the performance of the model heavily.

Decision Tree:

Confusion Matrix :

|  |  |  |
| --- | --- | --- |
| Actual/  Predicted | 0 | 1 |
| 0 | 22521 | 8263 |
| 1 | 48 | 4985 |

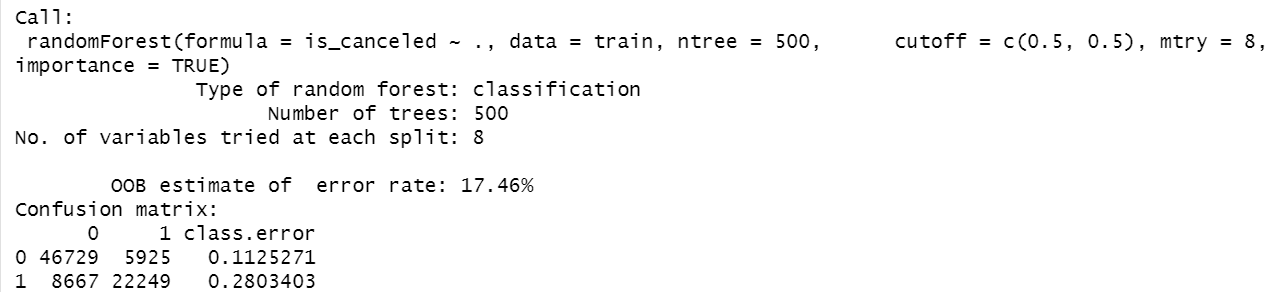
Accuracy = 76.79%

Random Forest

Confusion Matrix :

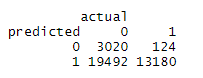
|  |  |  |
| --- | --- | --- |
| Actual/  Predicted | 0 | 1 |
| 0 | 46729 | 5925 |
| 1 | 8667 | 22249 |

Accuracy: 82.53%



The model shows a good performance. The OOB estimate of error rate is 17.46%, which means 82.54% of OOB samples were correctly classified by the random forest. 46729 hotels were correctly labeled as “No cancel”, and 22249 hotels were correctly labeled as “cancel”. The dataset contains 119391 rows of data, and the ntree=500, which is a huge amount of data for the model to make an accurate classification.

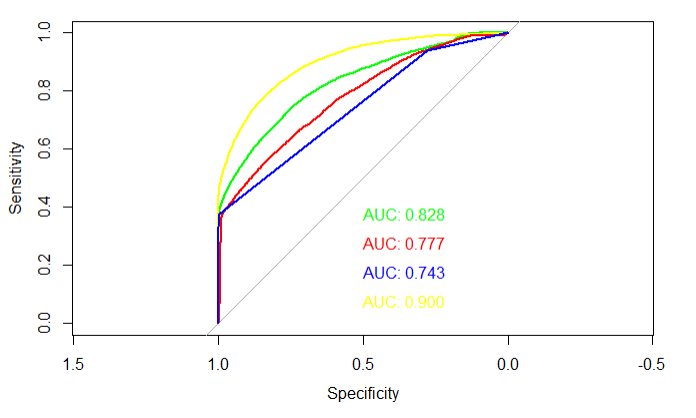
Naive Bayes



Accuracy of the model: 45.2%

It is obvious that Naive Bayes has poor performance. Mainly because of the high false-positive instances it predicted. Compared to Logistic Regression and Decision Tree, who have high false negatives, Naive Bayes exhibited a rather different result. The hotel should then make trade-offs when choosing which model to deploy since they all have pros and cons.

ROC curve for the models:



**Deployment**

We can see from the results that random forest achieves the highest auc, therefore, providing most useful information for us. We can see that lead time, deposit type, adr, previous cancellations, and stay cost total are highly influential in determining if a customer will cancel the reservation this time. The model shows us tendencies of customers so that hotels can make reasonable arrangements to avoid risks.

Accuracy of the model is still a problem. Even though random forest is our best model so far, it still only achieves 83% accuracy. On a large scale, the hotel will turn 17% of its business chance away. Since random forest is an ensemble method, we can incorporate more weak learners to enhance the performance of the classifier and achieve better performance.

Another aspect of our model performance is that it yields an equal amount of false-positive and false-negatives. False-positives will lead to losing a client because the room might be given to another client since the model predicted cancellation. False-negative will lead to underbooking and hence result in lost revenue.

The result may fluctuate in the short term, but in the long run it will appear to be 83% accurate, which is a major problem for the hotel to use this model to predict.

**Appendix**

Exhibit A: Data Descriptions: Variables

**hotel** (H1 = Resort Hotel or H2 = City Hotel)

**Is\_canceled** Value indicating if the booking was canceled (1) or not (0)

**Lead\_time** No. of days that elapsed between the entering date of the booking and the arrival date

**Arrival\_date** Arrival date, year, month and day provided in separate columns

**Stays\_in\_weekend\_nights** Number of weekend nights (Saturday or Sunday)

**Stays\_in\_week\_nights** Number of weeknights (Monday to Friday)

**Adults,** **Children, Babies** Number of each kind

**mealType** breakfast, lunch, and dinner

**Market\_segment, distribution\_channel,distribution\_channel** TA-Travel Agents TO-Tour Operators

**Is\_repeated\_guest**  repeated guest (1) or not (0)

**Previous\_cancellations** Number of previous bookings that were canceled **previous\_bookings\_not\_canceled** no of previous bookings not canceled

**reserved\_room\_type, assigned\_room\_type** Code of room type reserved

**Booking\_changes** Number of changes/amendments made to the booking

**Deposit\_type** Indication on if the customer made a deposit to guarantee the booking

**agentID, companyID** details of the travel agency that made the booking

**Days\_in\_waiting\_list** Number of days the booking was in the waiting list before confirmed **customer\_type** Contract, Group, Transient, Transient-party

**Adr** Average Daily Rat (divide sum of all lodging transactions by the total number of nights)

**Required\_car\_parking\_spaces**,**Total\_of\_special\_requests** special requests by the customer **Reservation\_status** Canceled, Check-Out , No-Show

**Reservation\_status\_date** Date at which the last status was set.

Exhibit B: Output of Logistic Regression

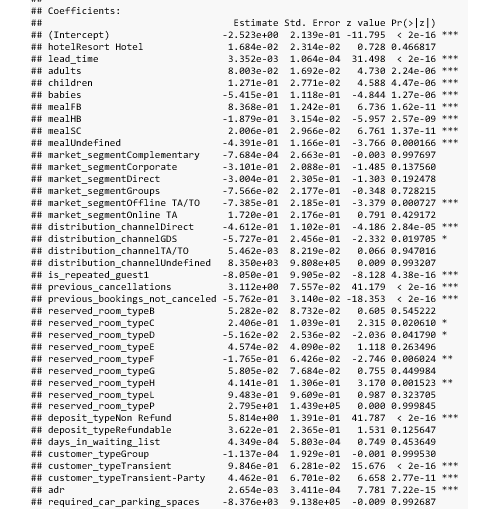


Exhibit C: Output of Decision Tree

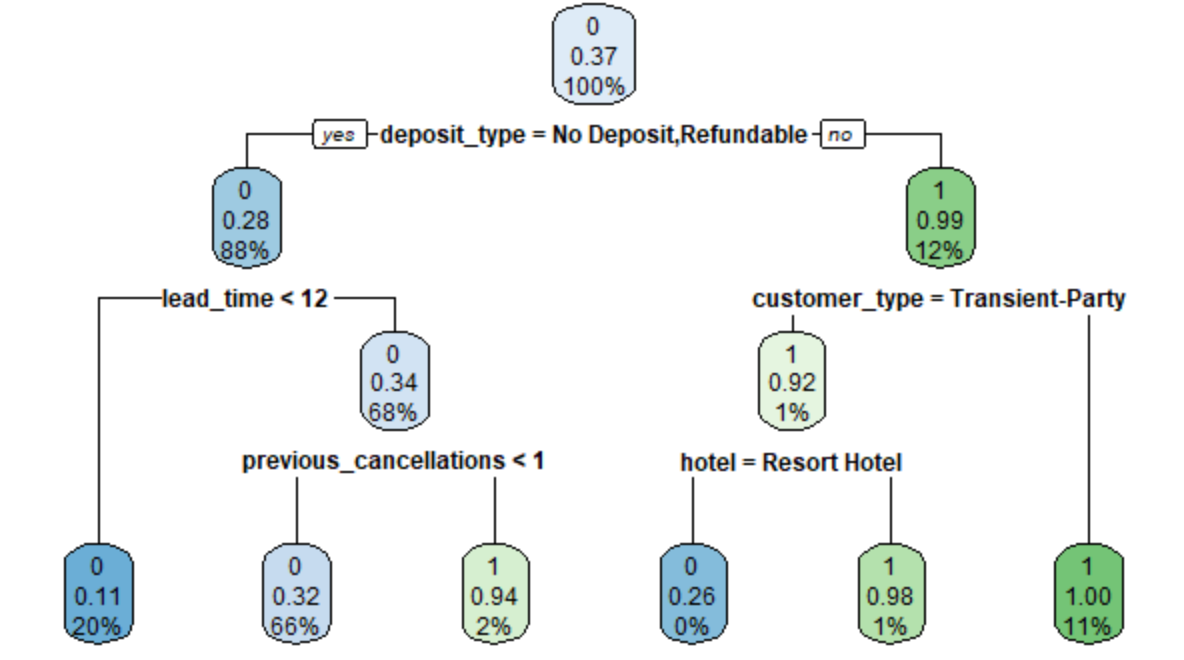
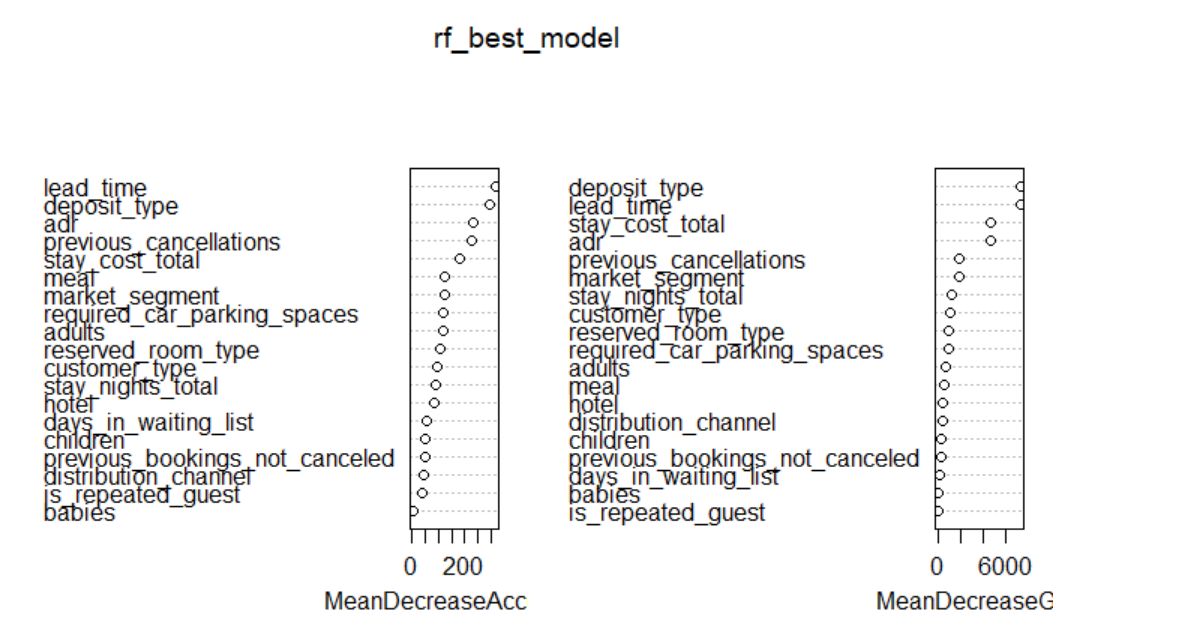


Exhibit D: Mean Decrease Accuracy Plot of Random Forest



**Reference**

* Antonio, N; Almeida A; Nunes Luis (2017). ‘Predicting hotel booking cancellations to decrease uncertainty and increase revenue’. Tourism & Management Studies, 13(2), 2017, 25-39
* Hotel Booking Demand. Retrieved from [www.kaggle.com/jessemostipak/hotel-booking-demand](https://www.kaggle.com/jessemostipak/hotel-booking-demand)