Case Study

In this section, be as detailed and specific about your ideas as necessary without deviating from the main theme of the case study. Some information required may not have been provided deliberately and thus require some assumptions on your part. In that case, please clearly state your assumptions in the response.

As an airline it is very important to know why our customers are travelling with us. For example, we can adjust our schedule (frequency of flights or timing of flights) to better satisfy our customer's needs and wants. It is particularly important to know whether a trip is taken for business or leisure purpose. Unfortunately, the purpose of a trip is often hidden and cannot be observed directly.

To address this issue, Hogwarts Air – a fictional airline – wants to build a model which can predict the trip purpose as a function of trip attributes. Hogwarts Air has collected training data by surveying several of their customers who have recently flown with their airline and asked the passengers to specify the intent of the trip – whether it was business or leisure.

The training data is stored in "trip_label_training.csv". Each row of the training data contains various attributes of a trip along with a *trip_label* (either B or L). A detailed list of trip attributes and their definitions are specified in the table below.

Using the data, please answer the following questions:

1. Build a prediction model and describe your model. Specifically,
   (a) Which features did you use?

- I've used the all the variables except - "*trip_date_outbound*", "*trip_date_inbound*", "*trip_od_outboard*", "*trip_od_inbound*", "*trip_od_inbound*", "*booking_city*" and "*booking_date*".
- These variables upon a quick time-series visualization did not convey a strong relationship to the target variable.

   (b) Which prediction model did you use? And why did you use it?

- I used "**Random Forest**" for building the model. Random Forest is one of the most efficient ensemble model which can be used for both regression and classification tasks.
- Since the behavior of passenger's travel data has a lot of non-linearity and variability, the Random Forest is a safe bet in predicting the target variable.

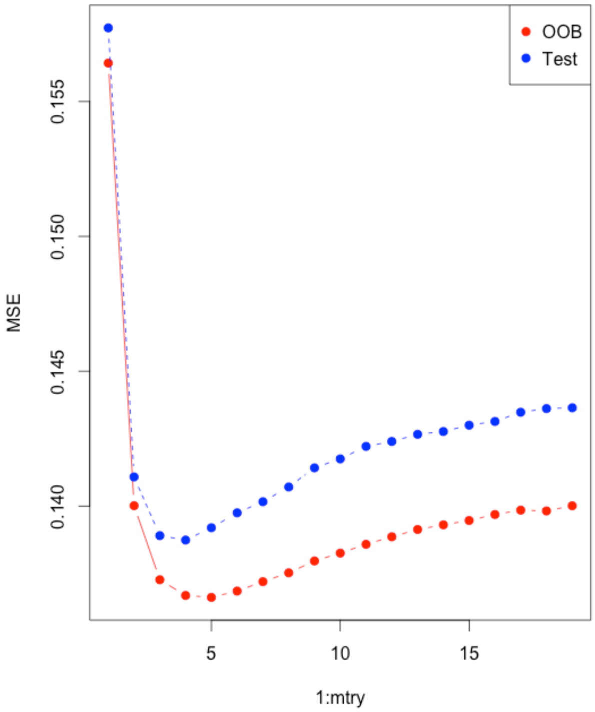
- Often the Random Forest creates a huge number of tress for a given set of variables and finally averages the final predicted values, there by prevents the over fit issues.
- Even tough the GBM's (Gradient Boosting Machine) were a good choice for the prediction accuracy, most of the cases the prediction results were almost similar to Random Forest.
- Moreover, keeping the computation cost and ease of tuning the model in mind, I finally choose Random Forest for the prediction.
- Didn’t preferred regression methods because they are more sensitive to the outliers and more suitable for the data which has more linear relationships.

(c) How did you select parameters (if any) for your prediction model?

- For the Random Forests, the best set of parameters can be configured by using the Out-Of-Bag error rate as a reference.
- The two main parameters for Random Forest are -

1. The count (*mtry*) of number of variables randomly sampled as candidates at each split while building the tree.
2. The number of trees to grow (*ntree*).

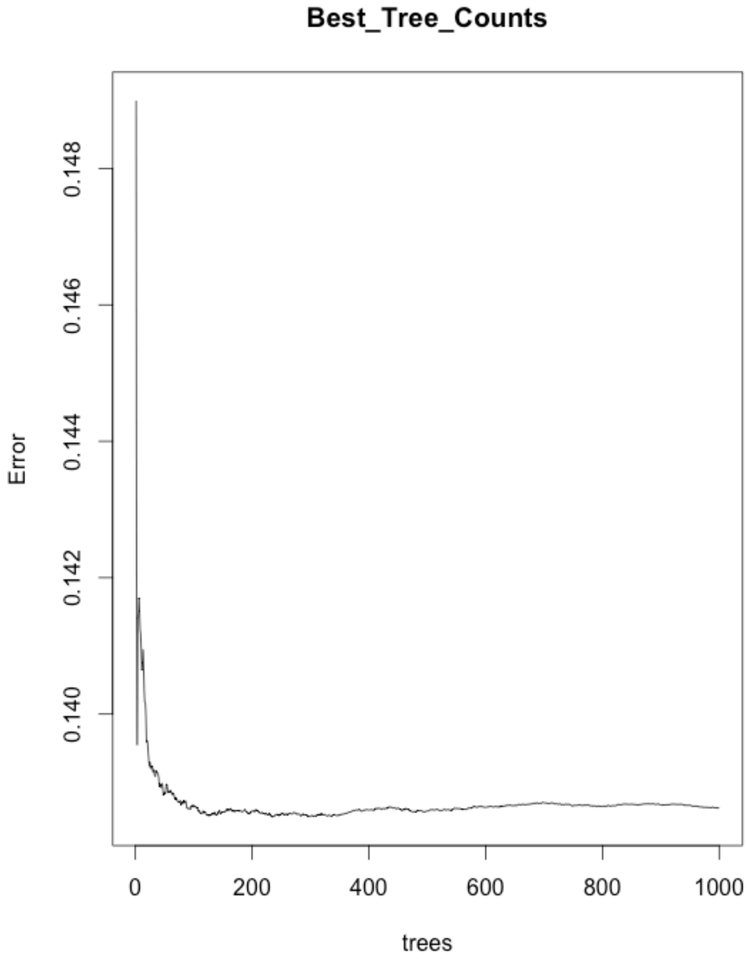
- **Model Tuning Result for “mtry”:**
  The optimum value for the “mtry” is as shown in the below plot:



Inference: Here from the above plot, the MSE for Out-Of-Bag error (OOB) and the Test error is significantly increasing from the value 4 and at the same time, minimum is at 4. So I've given "4" for the training model for "mtry" parameter.

- **Model Tuning Result for "ntree":**

  The best possible value for the number of tress can be seen from the below plot –

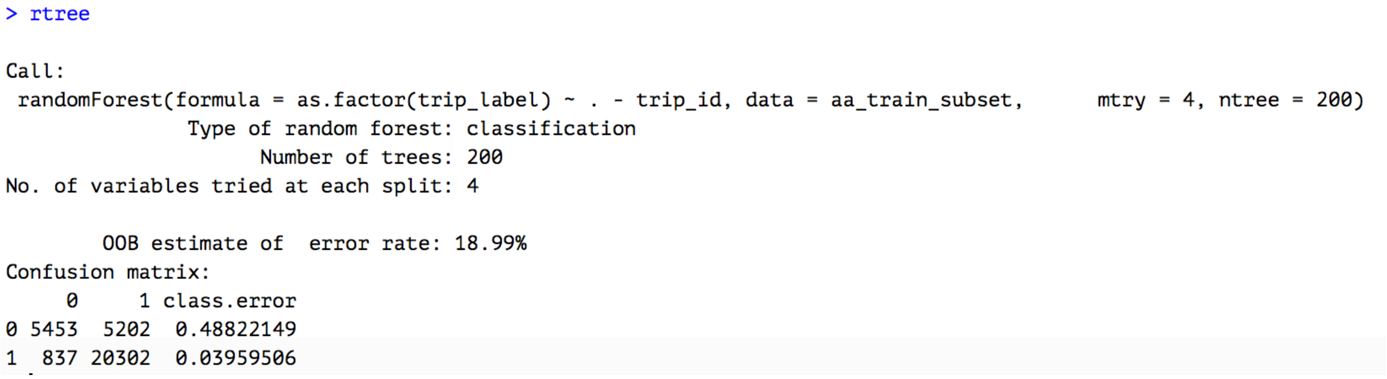


Inference: This plot depicts how Random Forest can be resistant to over fit. Even tough the number of tress increased over the time, the MSE error didn't changed. So choosing the number "200" for building the number of trees.

(d) Report the performance of the model. How did you evaluate the performance your model?

- The best way to evaluate the performance of the classification model can be done by using confusion matrix.
- From the confusion matrix, Out of Bag error estimate(misclassification) rate can be calculated by – (False negative+ False positive)/ (Total number of observations).

With the above formula, the model produced – 18.99% error rate.

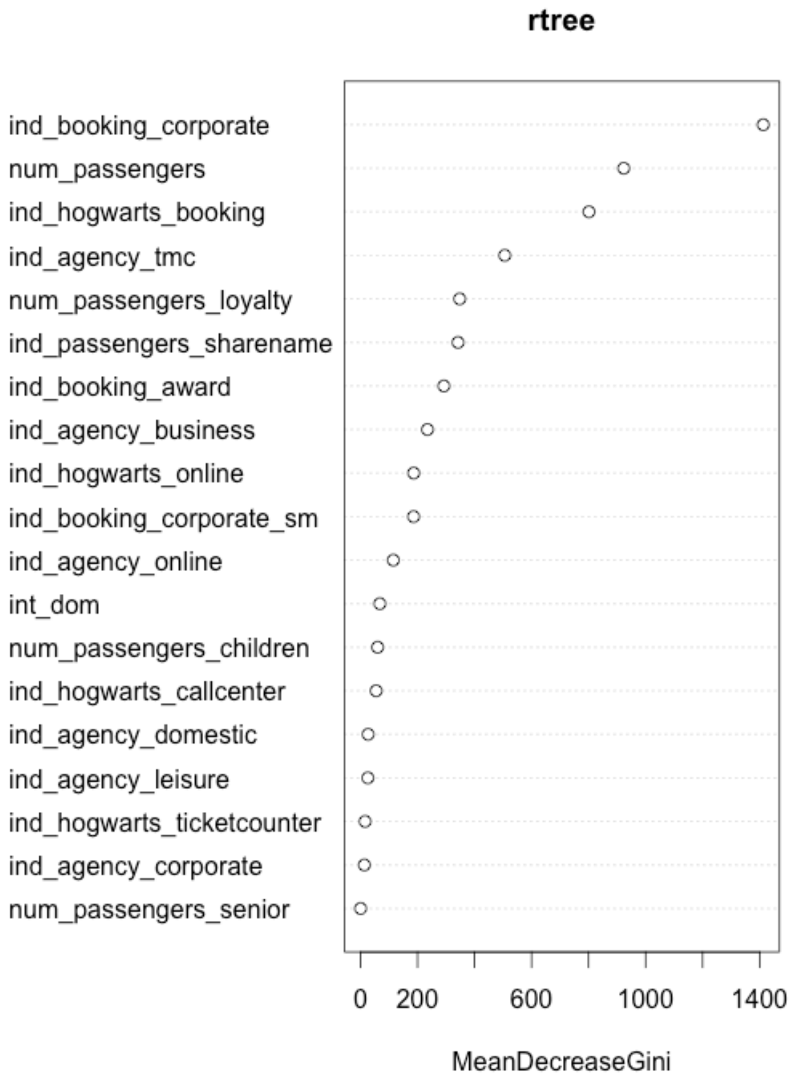


```
> rtree

Call:
 randomForest(formula = as.factor(trip_label) ~ . - trip_id, data = aa_train_subset,      mtry = 4, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 18.99%
Confusion matrix:
     0     1 class.error
0 5453  5202  0.48822149
1  837 20302  0.03959506
```

2. Based on your model, which of the trip attributes are main drivers of business/leisure trips?



- The variables - **"ind_booking_corporate", "num_passengers","" "ind_hogwarts_booking", and "ind_agency_tmc"** were considered to be the best in predicting the target variables. However, the importance of other variables with there relevant quantitative measures (Gini Index) were shown below.

Variables	**Gini Index**
ind_booking_corporate	1412.758594
num_passengers	923.4415198
ind_hogwarts_booking	801.3709446
ind_agency_tmc	505.9632436
num_passengers_loyalty	347.810112
ind_passengers_sharename	342.0147684
ind_booking_award	292.4829245
ind_agency_business	234.8087222
ind_hogwarts_online	186.6329115
ind_booking_corporate_sm	186.2069019
ind_agency_online	115.0372137
int_dom	67.7715861
num_passengers_children	59.9975301
ind_hogwarts_callcenter	54.8737567
ind_agency_domestic	26.7080868
ind_agency_leisure	25.7047548
ind_hogwarts_ticketcounter	15.9903654
ind_agency_corporate	13.3755747
num_passengers_senior	0.3538741

- Kindly see the attached R script for further details.

3. Apply your model to predict trip purpose for trips with unknown *trip_label*. The data to be scored is in “trip_label_score.csv”. Send us back a .csv file (“score.csv”) with just two columns {*ID, trip_label*}.

- Please find from the attached.

Attribute	**Description**	**Note**
trip_id	Row ID	
trip_label	Trip Label, either B for business or L for leisure	target variable
trip_date_outbound	The date for the outbound (departing) trip (if available)	date
trip_date_inbound	The date for the inbound (return) trip (if available). A missing value could indicate a one-way booking	date
trip_od_outboard	The origin and destination cities for the outbound (departing) trip (if available)	categorical, string
trip_od_inbound	The origin and destination cities for the inbound (return) trip (if available). A missing value could indicate a one-way booking	categorical, string
num_passengers	Total number of passengers in the booking	integer
num_passengers_senior	Number of senior (age 65+) passengers	integer
num_passengers_loyalty	Number of loyalty program passengers	integer
num_passengers_children	Number of children	integer
ind_passengers_sharename	Whether some passengers share the same last names	binary
booking_city	Closest airport city to the customer's home address	categorical, string
booking_date	Date trip was booked	date
ind_domestic	Whether it is a domestic trip or international trip	binary
ind_booking_award	Whether the booking is through mileage redemption (i.e., free trip)	binary
ind_booking_corporate	Whether the booking is made with a corporate account	binary
ind_booking_corporate_sm	Whether the booking is made with a small-medium corporate account	binary
ind_agency_domestic	Whether the booking is made with a domestic travel agent	binary
ind_agency_leisure	Whether the booking is made with a travel agent mostly working in the leisure sector	binary
ind_agency_business	Whether the booking is made with a travel agent mostly working with business accounts	binary
ind_agency_corporate	Whether the booking is made with a travel agent working almost exclusively with corporate accounts	binary
ind_agency_online	Whether the booking is made through an online travel agent like Expedia or Hotwire	binary
ind_agency_tmc	Whether the booking is made with a Corporate Travel Management Company (TMC)	binary
ind_hogwarts_booking	Whether the booking is made through Hogwarts Air	binary
ind_hogwarts_online	Whether the booking is made through Hogwarts Air's website	binary
ind_hogwarts_callcenter	Whether the booking is made through a Hogwarts Air call center	binary
ind_hogwarts_ticketcounter	Whether the booking is made through a Hogwarts Air ticket counter	binary