**Data Science Challenge**

My work is in three stages: exploration, data preparation/wrangling and model building, with each stage encompassing the thought process and methodology used to accomplish it. The stages are as follows:

**Exploration**

This was done in part to answer the questions posed in this assignment and as well as to study the data. The summary of all the fields were derived and stored in the files numerical\_variables\_summary.csv and categorical\_variables\_summary.csv while variables that are all null were excluded.

Plots of the variables are presented in the folder plots, for all the data, for isFraud==False and isFraud==True to see the variables. The Histogram for the transactionAmount (Compact\_Histogram\_transactionAmount.jpg) where I found that the data and for isFraud==False follow a loglaplace distribution and for the isFraud==True follows weakly a pareto distribution.

**Data preparation**

With data preparation, I did many things: to remove duplicate transactions and add variables. I first added a field timestamp; by converting transactionDateTime to time stamp to help in ordering the transactions for each user and I also added the account age; this helped in adding other variables and in duplicate removal

Removing duplicates: this was done by ordering transactions per user using timestamp then eliminate duplicates from the data when changeable variables like timestamp, transactionDateTime, etc but leaving accountage (in days) because multiple swipes will occur in one day.

Add variables: other variables were added; transaction\_rate which the rate since account opening (total transaction amounts divided by number of days). I considered this important because it is easy to compute and banks use it (as Average Ledger) and a thief will likely use more than the owner by intuition. Transaction\_ave which is the average transactionAmount with respect to number of times of use, this is because the use by an owner will be uniform, still from intuition.

A plot of histograms of the transactionAmount, counts, reversal amounts and reversal counts was drawn in file Histogram\_transactionAmount.jpg for each user.

Reversals were also removed from the data.

**Model building**

I built four models: extreme boosted trees, Adaptive gradient boosted trees, decision tree and random forest, the results of each are in the folder models, with accuracies: 0.761057, 0.704934, 0.986417 and 0.999397 respectively. The confusion matrices and variable importance are also saved in the folder models.

The best model is random forest (rfmodel) and we are taking the best because the first two are poor and the third is basically part of the random forest and we should not do stacking since the best is almost perfect.

The top five most import variables were transactionAmount, cardLast4Digits, transaction\_ave, transaction\_rate and cardCVV. Among the category variable merchantCategoryCode, online\_ retail and online\_gifts are most importance while cable/phone and gym are least.

**Conclusion**

From the way random forest works, importance just means that when the variable was used, it improved accuracy by a large margin, which could be any number of times and as such we cannot point to value as the cut-off or specific properties of it as fraud characteristics but nevertheless we can try some.

Looking at cardLast4Digits and cardCVV being two of the most important variables, a plot showed peaks on some numbers which at first glance may be easily memorized numbers but we will expect them to follow a uniform distribution as the histogram portrayed, a further look into them will useful.

We could improve the results by looking into merchant name and classifying them into classes according to frequency of occurrence, merchant name was not included because there were over 2000 and creating dummies will be a nuisance. With more time I can look into merchant names to improve results and characteristics of important variables.

Null values were very few, with time I could look into imputing using high performance model if I have proof that more data will improve accuracy: this is done by plotting number of records used for training against accuracy, if the plot is increasing then that is the prove we need.