CS4035 - Cyber Data Analytics
Assignment 1

Prerak Mody - 4777042
Sujit Shankar Jaishankar - 4779657

May 2019

1 Visualization Task

We perform various vizualization tasks and plot them in the hope to better understand the data and explain our modelling results explained in Section3. The submitted Jupyter notebook file will further detail the visualizations made along with the data processsing process. In this part, we have provided graphs for certain features we found interesting after our initial analysis

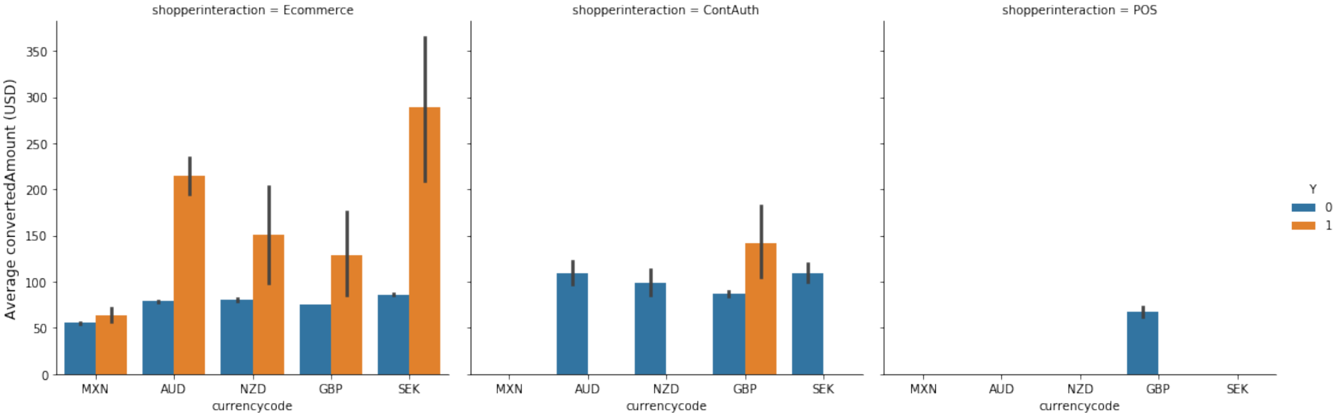

Figure 1: (Q1) - Histograms for Average Amount in USD based on currencycode and shopper interaction

Plot1 - In plots in fig. 1, the categorical variables - currencycode and shopperinteraction are visualized against the average amount in US Dollars. We notice that most of the fraud transactions happened in Ecommerce domain. Only few of them appeared in ContAuth(subscription) and none at all in POS. It is interesting to note that when using subscription services(ContAuth), fraudulent transactions were found only in GBP currency. Apart from that, we infer that majority of Ecommerce fraud transactions happened in SEK and AUD. In other word, there is a significant difference in fraud occurrence among different currencycodes. It implies that currencycode and shopperinteraction can be used as informative features.

Plot2 - A bulk of fraudulent transactions are carried out by three types of cards - mccredit, visaclassic and visadebit. The special cards like visabusiness and visasignature are less in number probably because the number of these cards issued is less. Also, they might have extra security layers that might prevent fraudsters from using them otherwise.

1.1 Inferences

Other inferences we made from our exploratory analysis are listed below.

1. No fraudulent transactions were made via PoS (Point of Sale).
2. No fraudulent transactions on Electron and Visa corporate cards.

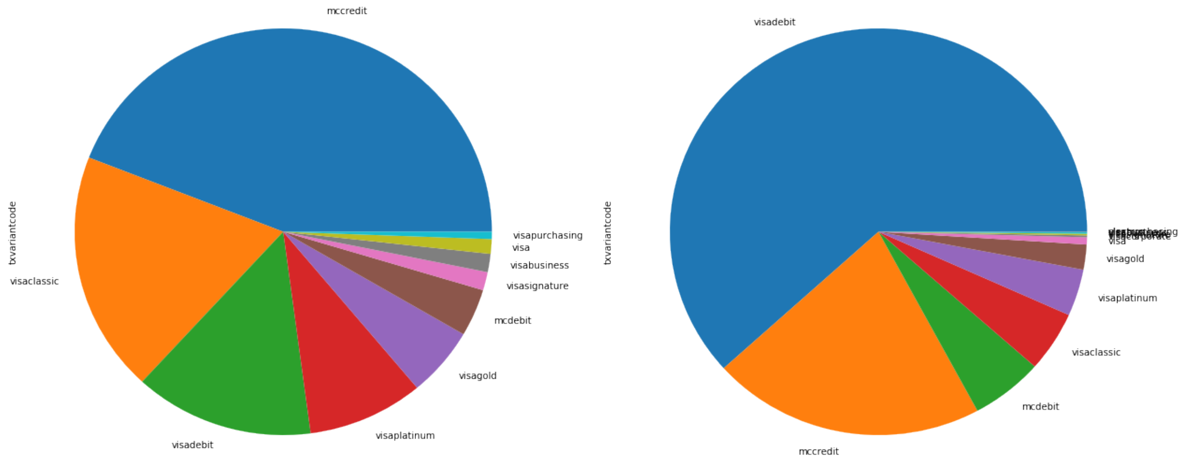

Figure 2: (Q1) - Pie Charts for the transactions with respect to txvariantcode (Fraudulent on the left, benign on the right)

3. In fraudulent cases, cvcResponseCode 2 and 3 are not present. That is, fraudulent transactions had CVC codes that were either unknown or a match.

2 Imbalance Task

We have provided the ROC curves for three classifiers – logistic regression classifier, random forest classifier and the adaboost classifier :-

- Logistic Regression - This classifier outputs ROC curves (fig. 3 and fig. 4) which look almost similar for the SMOTE and unSMOTEd datasets. The area under the curves (AUC) values are also quite close. The only difference can be noted in the top left corner of the curves where the unSMOTEd dataset has a slight depression which signals the loss of true positive rate for an equivalent false positive rate.
- Random Forest - This classifier outputs ROC curves which follow a similiar pattern but produce AUC values with a considerable gap. Here we can clearly see the advantages of training a classifier on SMOTEd data.
- AdaBoost - This classifier also has a similiar pattern in the SMOTE and unSMOTEd datasets.

It is clear from some of the graphs (i.e RandomForest's ROC Curve) that applying SMOTE is a good strategy for the given dataset. Unfortunately, it is still not possible to see its clear advantages by looking only at the ROC curves as the true positive rates of unSMOTE'd data are quite low. On using SMOTE, the true positive rates improve drastically and hence using can be considered good practise.

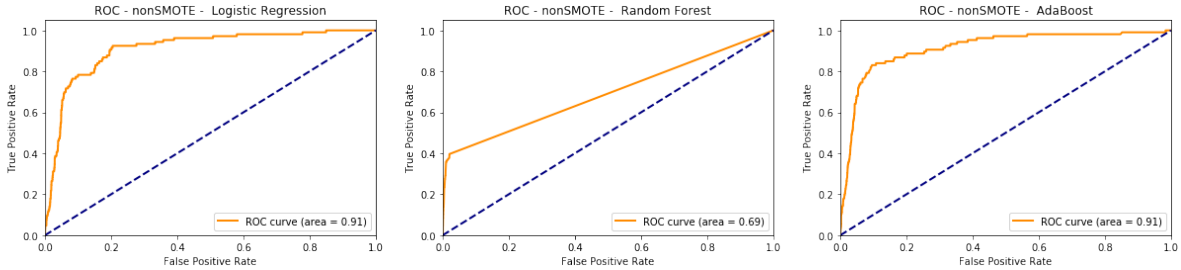

Figure 3: (Q2) - ROC Curves for unSMOTEd data

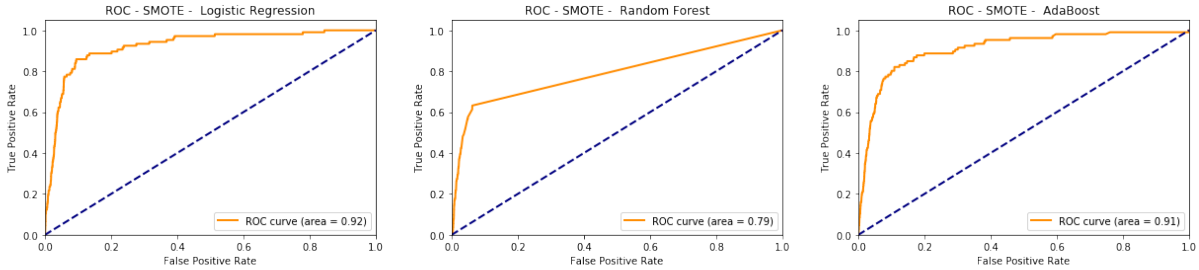

Figure 4: (Q2) - ROC Curves for SMOTEd data

3 Classification Task

3.1 Data Pre-processing

Since the provided dataset consisted of a mix of categorical and numerical variables, with the categorical variables having a high amount of unique values, we performed the following data preprocessing steps :-

- First, we convert the column - *simple_journal* to Y which is 0 for non-fraudulent cases and 1 for fraudulent cases.
- The column - *amount* contains values in the local currencies and thus we convert them to US Dollars and finally perform min-max normalization
- We drop other unimportant columns such as *txid*, bin, date columns (creationdate, *bookingdate*) and personal identifier columns (mail_id , *ip_id*, *card_id*)
- We convert categorical columns such as *txvariantcode*, *currencycode*, *shopperinteraction*, *cvcresponsecode* and *accountcode* to numerical columns
- We one-hot encode the top fraud values for *issuercountrycode* and *shoppercountrycode*, specifically for [MX, AU, GB, SE, NZ amongst others]

3.2 Black Box Algorithm

We use an ensemble of classifiers to leverage the best of each. They are as follows :-

- Logistic Regression (with regularization of 0.1) - This classifier usually tends to lean more towards a higher true positive count.
- AdaBoost (with 250 decision stumps and a learning rate of 1.0) - This classifier usually tends to lean more towards a conservative true positive count, but comparitively lower false positive amount
- XGBoost (subsample= 0.5) - The performance of this classifier lies in between the ones mentioned above and serves as a good balance between them.

This gives us metrics as can be seen in fig. 5

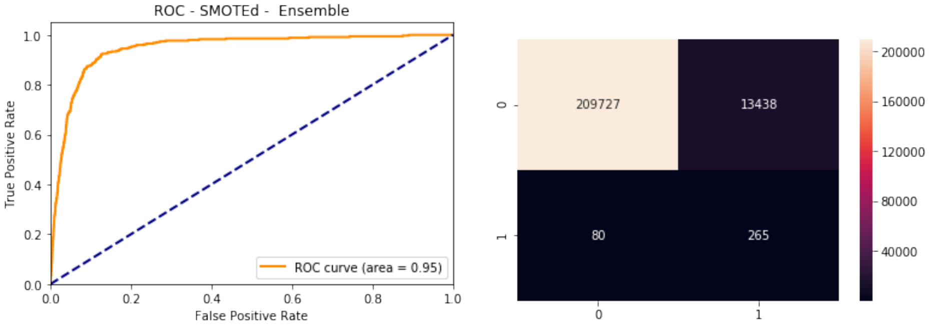

Figure 5: (Q3) - ROC Curve for Black Box Algorithm

3.3 White Box Algorithm

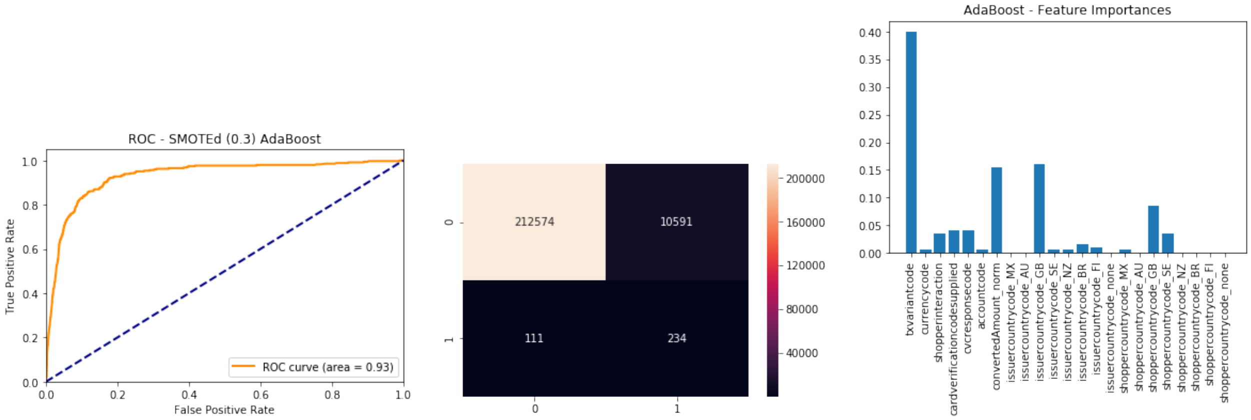

Figure 6: (Q3) - Performance Metrics and Model Information for the White Box Algorithm - AdaBoost Classifier

If we look at the the third plot in fig. 6, we can see that various features have their importances plotted in a bar chart. This reflects some of the results that we see in our data visualization task such as :-

- The *txvariantcode* which represents the type of the card being used was seen to be an important feature in fig. 2 as there was a different distribution of card types between fraud and non-fraud data.
- The *convertedAmount_norm* column is a min-max normalized column of the US Dollar converted column containing information on the transaction amount. This can be seen in fig. 7 as well where the average value of fraud vs non-fraud transactions is clearly different
- The *issuercounttrycode_GB* and shoppercountrycode_GB columns are also important discriminators as can be seen fig. 1 where certain types of shopper interactions (i.e "ContAuth") have fraud transactions only in the GBP currency.

Appendix

Installation Requirements

To run the code submitted with this report, follow the given instructions step-by-step

- Ensure that you have Python2.7 (or greater) installed
- Open a terminal, change directory to the submitted folder and run *pip install -r requirements.txt*
- In the same terminal, run *jupyter notebook* which opens a local server in your browser.
- Click on *Assignment_Final.ipynb* and follow the instructions within that.

Task A - Data Visualization

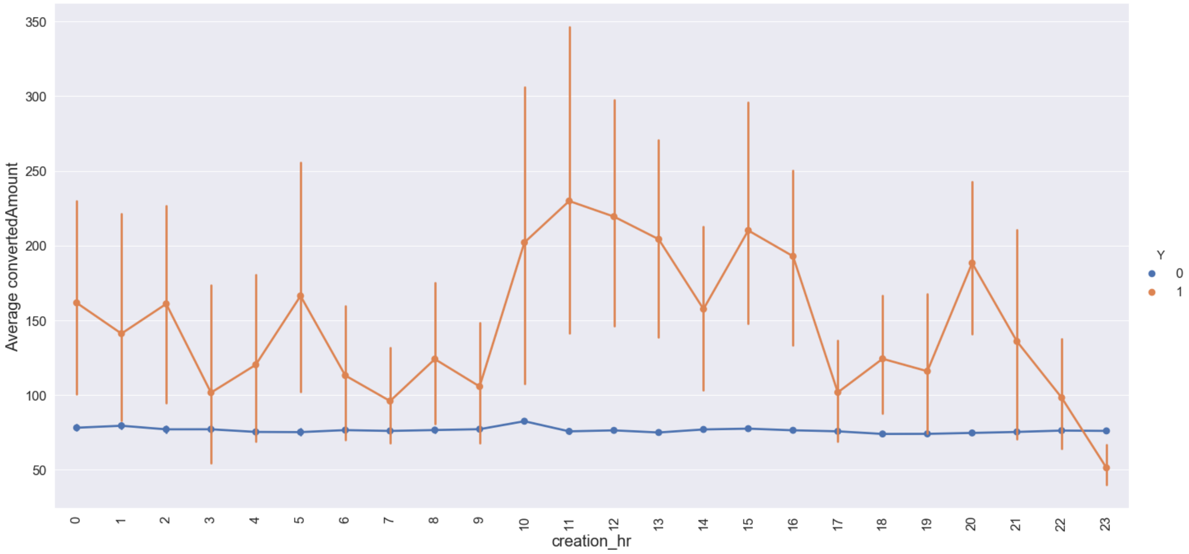

Figure 7: (Q1) - Amount for both fraud(1) and benign(0) transactions over time.