Deliverable II

Literature Study and Exploratory Data Analysis

There are many studies done on anomaly detection with IDA, LDA and PCA. Multiple models with good accuracies based on ANN, Logistic regression and Random forests are already present on the given problem.

http://www.questjournals.org/jrhss/papers/vol8-issue2/B08020411.pdf

https://ieeexplore.ieee.org/document/8123782

https://pdfs.semanticscholar.org/0419/c275f05841d87ab9a4c9767a4f997b61a50e.pdf

In this study, Credit card fraud detection is a typical uncertain domain, where potential fraud incidents must be detected in real-time and tagged before the transaction has been accepted or denied. The inclusion of uncertainty aspects impacts all levels of the architecture and logic of an event processing engine. This enables the implementation of event-driven applications possessing uncertainty aspects from different domains in a generic manner. the preliminary results are encouraging, showing potential benefits that stem from incorporating uncertainty aspects to the domain of credit card fraud. In the other study, the research investigates that machine learning like Naïve Bayes, Logistic regression, Random forest with boosting and shows that it proves accurate in deducting fraudulent transactions and minimizing the number of false alerts. Supervised learning algorithms are novel ones in this literature in terms of the application domain. If these algorithms are applied to the bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions. And a series of anti-fraud strategies can be adopted to prevent banks from great losses and reduce risks. By comparing all the three methods, they found that a random forest classifier with a boosting technique is better than the logistic regression and naïve Bayes methods. Also, one of the other sides, the researcher implemented the fraud detection technique used by VISA and MasterCard. Also, the study shows up that the Neural network is the latest technique that is being used in different areas due to its powerful capabilities of learning and predicting. Also, the study tries to use this capability of neural network in the area of credit card fraud detection as we know that Backpropagation Network is the most popular learning algorithm to train the neural network so in this paper BPN is used for training purpose and then to choose that parameter that plays an important role to perform neural network as accurately as possible.

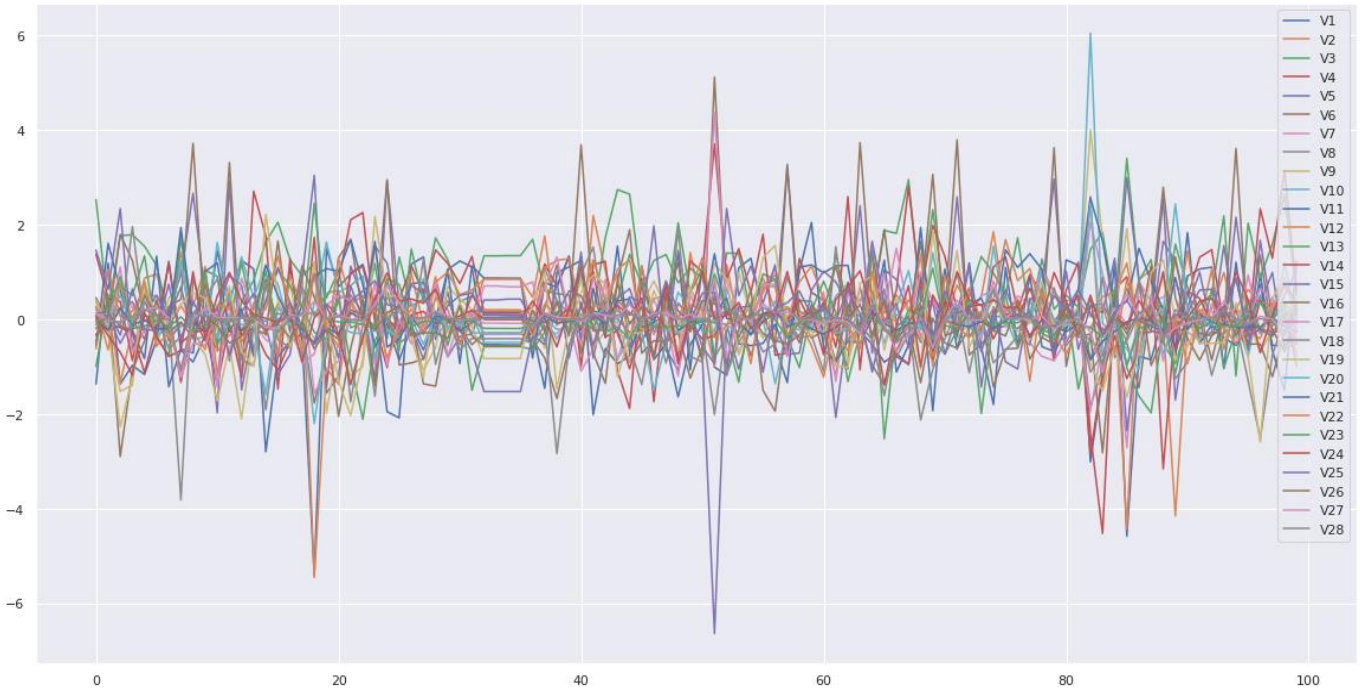
Exploratory data analysis

Since nearly all predictors have been unidentified, I decided to focus on the non-unspecified predictor's time and amount of the transaction during my EDA. The distribution of the financial value of all transactions is heavily unbalanced. Most transactions are relatively small, and only a tiny fraction of transactions come even close to the maximum amount.

Observing the data summary

- ✓ Prepared Data: A process to gather context to the input data. Understanding the data for preprocessing and cleaning of datasets.
- ✓ The two columns “amount” and “time” were not normalized. The remaining columns were normalized using Principal Component analysis.
- ✓ Oversampling (Using SMOTE): The fraud transactions are 492 samples which is unbalanced.
- ✓ Training and Testing Subset: As the dataset is imbalanced, many classifiers show bias for majority classes.

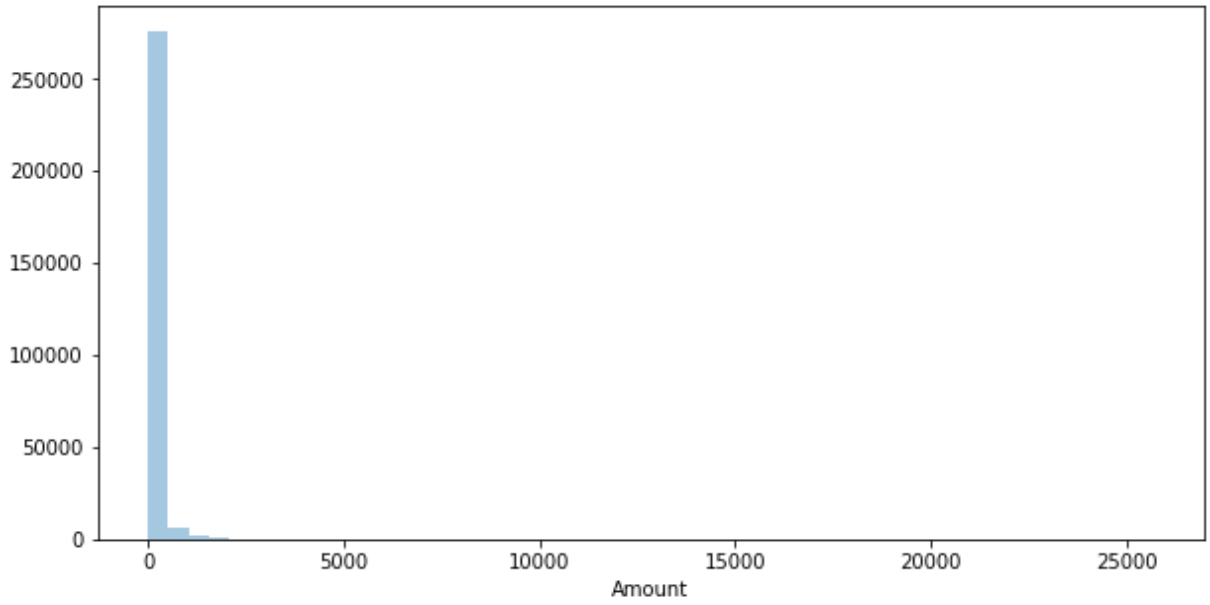
1. Observing features through line plots



The above line chart was constructed to observe the anonymous features in the data

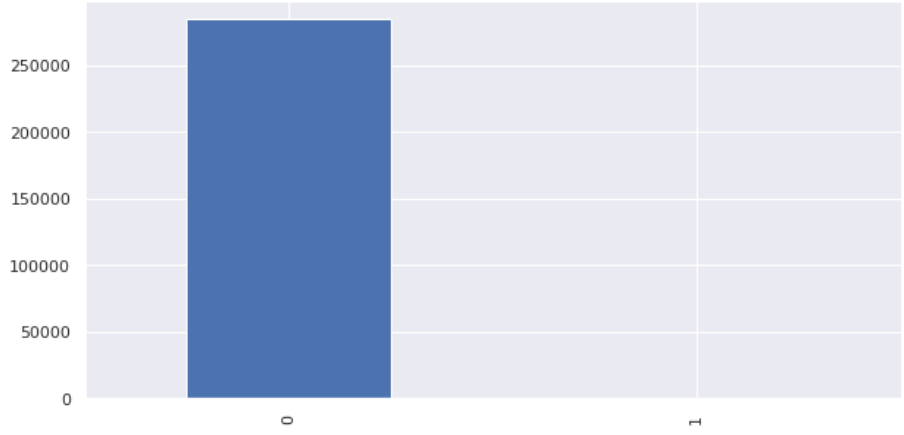
2. Observing distribution of amount across transactions

It can be observed from the below histogram plot of the amount that most of the amount transacted lies between 0 and 5000 even though other entries are present, they are very low in the count.

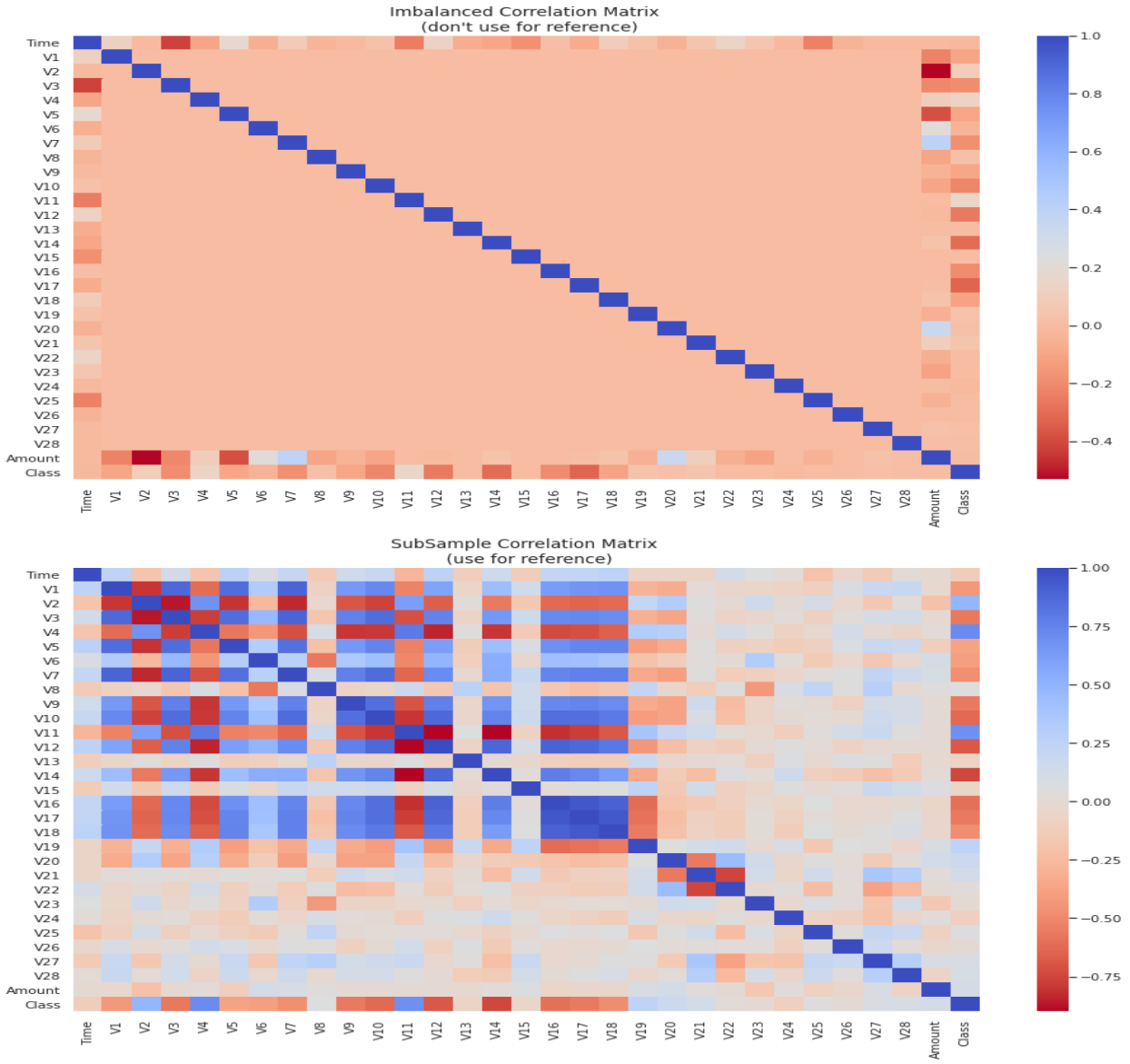


3. Observing the nature of the predicted class

The below histogram of the class predicted variable shows a very high imbalance in the predicted value. Class 0 (Normal transaction) has more than 250000 values, as compared to 492 Class 1(Fraudulent transactions).



4. Observing the correlation of anonymous input features



The above figure shows that the anonymous components in the dataset are highly uncorrelated and there is a good chance of them coming from the result of a PCA analysis.

5. Observing the nature of the predicted balanced class

To create our balanced training data set, I took all the fraudulent transactions in our data set and counted them. Then, I randomly selected the same number of non-fraudulent transactions and concatenated the two. After shuffling this newly created data set, I decided to output the class distributions once more to visualize the difference.

Outcomes

- Exploratory data analysis was performed with multiple plots on the relation of input features and their distribution.

- Data preprocessing was performed with Pandas

- Logistic regression, random forests and CNNs will implemented to predict credit card fraud detection dataset.