CREDIT CARD FRAUD DETECTION

DSC550 FINAL DOCUMENTATION

Ramizuddin Mohammed Shabuddin
rmohammedShabuddin@my365.bellevue.edu

Bellevue University

Abstract (Problem):

Credit card fraud is one of the major problems in financial services. This will cost billions of dollars every year. The United States is the most credit card fraud-prone country in the world. During the pandemic, the number of credit transactions and online shopping has increased, and this results in an increase in the number of credit card fraud activities. Fraudsters always work to find new means to conduct fraud in credit card transactions. Losses to financial institutions can be avoided by detecting credit card fraud and alerting banks about potentially fraudulent transactions. With the available credit card fraud dataset, I will perform some graph analysis to understand the dataset. There could be false positive or false negative fraud detection. We may need to perform scaling on the dataset. I will be using machine learning algorithms to detect fraudulent transactions. Companies like Visa are looking to identify new solutions using artificial intelligence to solve credit card fraud. Having a fraud transaction will create a business impact too. Because the customer will be losing the trust in the financial institution. So, credit card companies need to recognize fraud transitions to keep the customers not charged for wrong items and keep them happy. For my analysis the dataset from the Kaggle.

Kaggle link: https://www.kaggle.com/mlg-ulb/creditcardfraud

Approach:

The Kaggle dataset contains transactions that occurred in 2 days in the month of September 2013. There were 492 fraud transactions out of 284,807.
Dataset consists of below information:

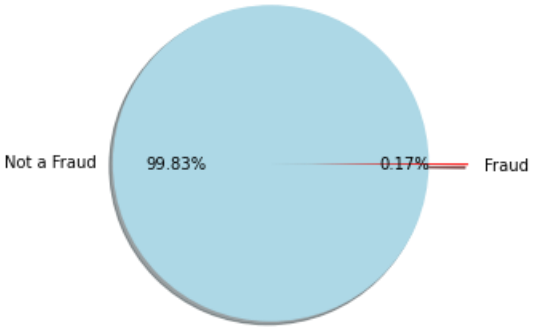
- Columns V1 to V28 are result of Principal Component Analysis (PCA) which has been done due to data compliance.
- Amount is the Transaction amount
- Class 0 indicates non-fraud and 1 indicates fraud
- Time is the time elapsed between each transaction.

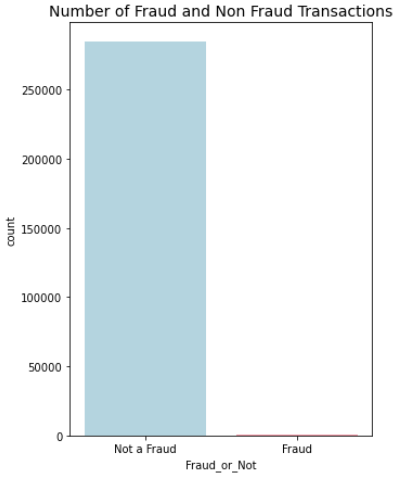
The data set contains principal components obtained with PCA. Using the above dataset, I will perform some graph analysis.

Graphical Analysis

Fraud vs Not a Fraud Transactions

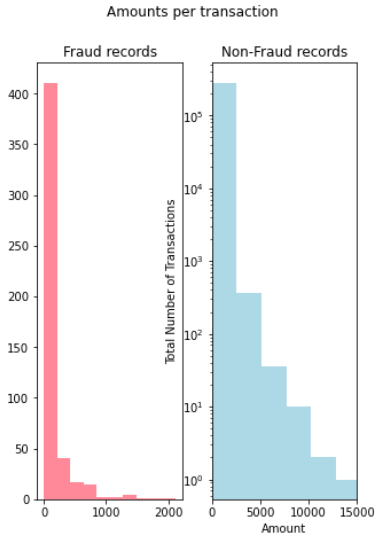
As outlined in the Chart below 0.17% percent of data is Fraud transactions. The dataset is highly imbalanced. Since the principal components are PCA transformed, scaling is performed to normalize the data as it could impact the performance of the model and standard scaler method is used.





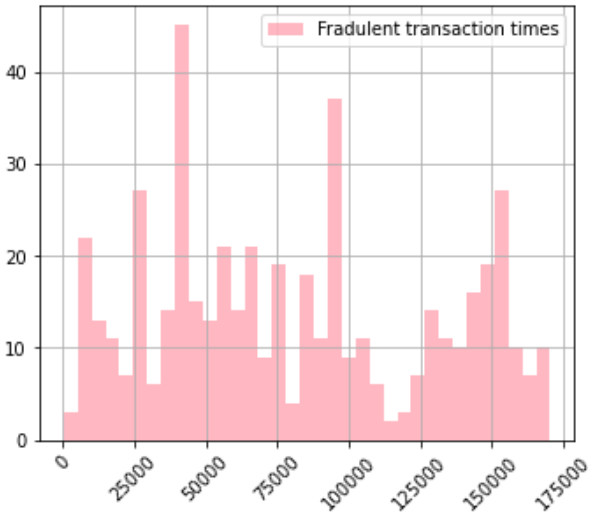
Total number of Transactions vs amount based:

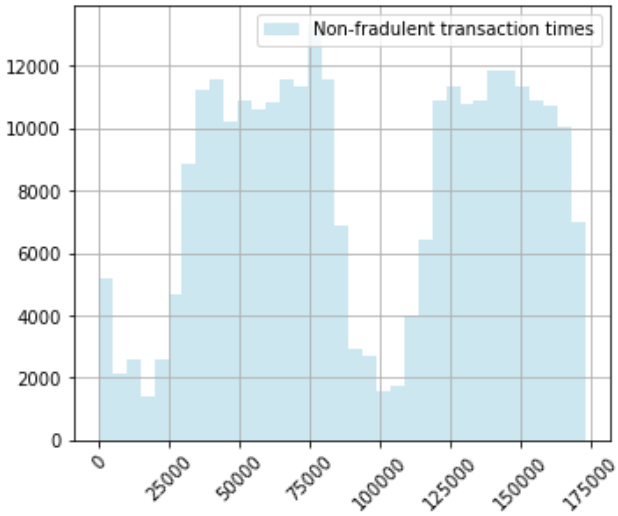
The proportion of fraud transactions are in hundreds, whereas the non-Fraud records are in 10000. Which is outlined below.



The transaction times:

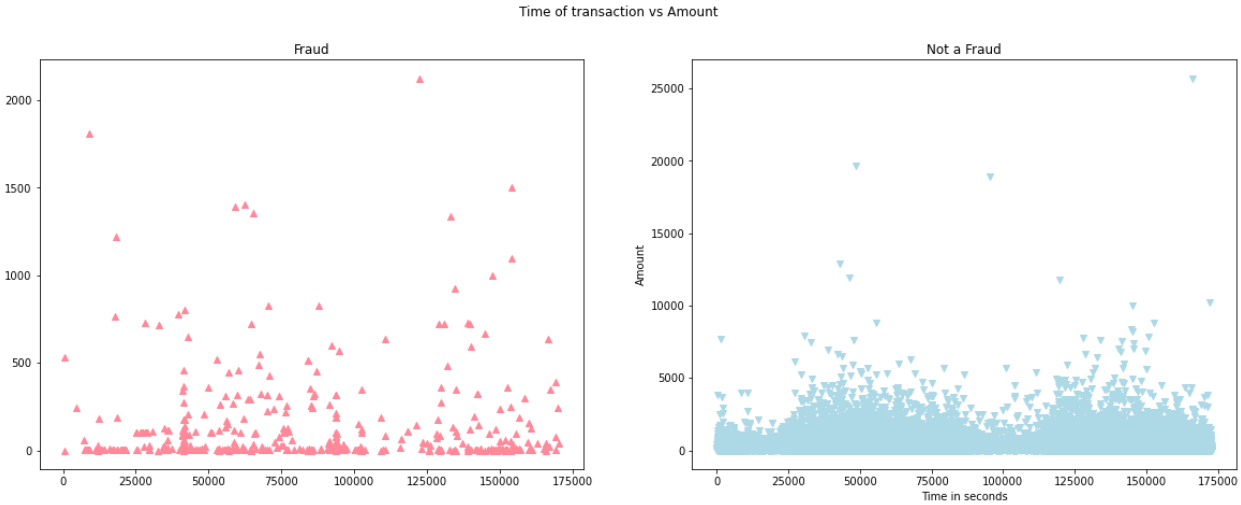
The Histogram has transaction time for fraud and non-fraud transactions over the transaction times. If we notice the graph below, we see some peaks times in fraudulent transactions times. But this cannot be an effective feature for the model.





Transactions time vs amount:

The scatter plot below has information about the fraud and non-fraud transactions over the transactions time. As these cannot be a feature for the model.

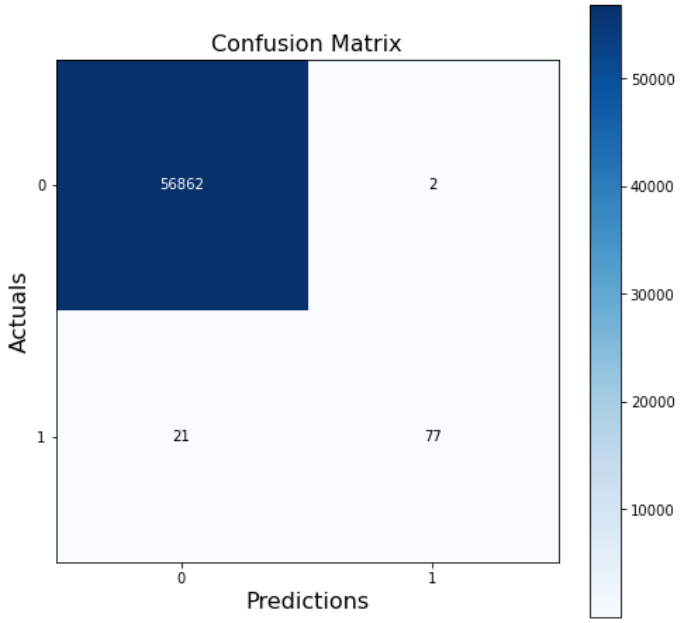


Feature Selection

Using the feature selection technique, we can verify the above. Variance threshold is used to select the feature with variance above a threshold. Variance threshold is calculated based on probability density function of a particular distribution. When a feature has 95% or more variability, then this is very close to zero and the feature may not help in predicting the model. This can be removed. The values with True are the features selected using Variance threshold technique. The columns from V23 to V28 and 30 are removed. Then, the data is split into 80% of train dataset and 20% of test dataset.

Model Selection

The AUC-ROC (Area Under the Curve – Receiver Operating Characteristic) curve is a performance measurement for the classification problems at various thresholds. Area under the Curve is one of the good metrics to evaluate the score of classifiers. It validated the method identifies if the transaction is a risk or not. Cross validation ROC-AUC scores are identified for different models. The Random Forest classifier score is 95.14. The model is evaluated using the test dataset and AUROC is 95.83 The results can be evaluated by confusion matrix. The confusion matrix helps in identifying if the predictions were correct or not. With higher the values of confusion matrix in the diagonal where predictions are right the better would be the model.



Conclusion:

As outlined in the abstract the credit card fraud is one of the issues where financial institutions have gone through many losses. These institutions must identify those fraud and minimize those loses. The AUROC metrics for Random Forest classifier is 95.83%. This will help in identifying the frauds. The log loss is helpful in identifying the better performance of the models in future datasets. The log loss is another metrics used and the score of 0.0139. This model can be used for predicting fraud transactions.