Credit Card Fraud Detection – DSC550 Final documentation

By: Santosh Kumar Omprakash

Bellevue University
1000 Galvin Rd S, Bellevue, NE 68005
somprakash@my365.bellevue.edu

Problem Statement:

Financial fraud is a major concern in banking and financial services and costs millions of dollars every year. Credit card fraud continues to rise due to highly imbalanced dataset and non-stationary nature of data. It is estimated that some organization have losses up to 5% due to fraud and recent studies point that fraud activities have risen and expected to increase in future. Losses to financial institution can be avoided by detecting credit card fraud and alerting banks about potential fraudulent transactions.

Proposal:

The first step in building a model is to get the dataset. One of the main issues with credit card fraud detection is unavailability of real dataset. Since the credit card data will have sensitive details about the customer, there are not many datasets available due to privacy issues. Another challenge with credit card dataset is that it is highly imbalanced where there are more legitimate transactions and few fraudulent transactions. Millions of transactions are processed every day and the size of the dataset will be huge. Feature selection technique is applied to use only required features and machine learning algorithm using Random Forest Classifier is used to evaluate and effectively predict fraudulent transactions.

Implementations:

Accuracy is not a suitable metrics for all machine learning algorithms. Specifically, accuracy cannot be used as a metric for credit card fraud detection due to imbalance in dataset. The cost of error in misclassifying fraudulent occurrence is more than that of misclassifying legitimate occurrence. To outline the steps followed, graph analysis is performed to understand some trends in the dataset. As the dataset has many principal component elements, feature selection techniques are used to determine the best features that would contribute to the model performance. In machine learning, classification of data into categories is needed for observation. Based on the input values classification, the dataset is split in train and test for testing and evaluation of the model. Cross validation methods are then applied to obtain average score for the model. Finally, the test data is used to validate the model and plot the confusion matrix.

The dataset that is selected has transactions from European cardholders made in 2013. It has 285,000 transactions out of which 492 are fraudulent. Due to privacy concerns, some principal components are PCA transformed. Time and Amount values are not transformed. Class value 0 is non-fraud and 1 is fraudulent transactions. The dataset is highly imbalanced as seen in figure 1. 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.

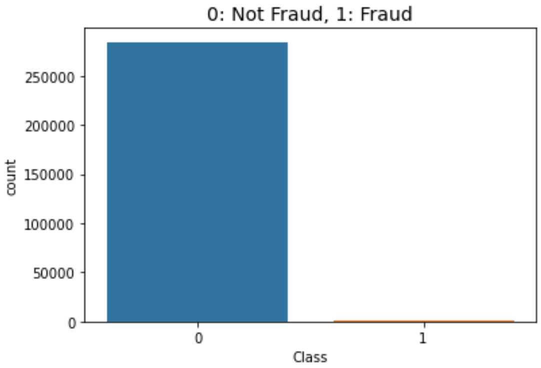

Figure 1: Fraud and non-fraud transactions

Identifying the best features is essential in predicting fraudulent transactions. Fraud detection is anomaly detection characterized by imbalance between classes and can be detrimental factor in feature selection techniques. Histogram plots in figure 2 has transaction times for fraudulent and legitimate transactions. Though there are some peak times in fraudulent transactions times, it cannot be used as effective feature for the model.

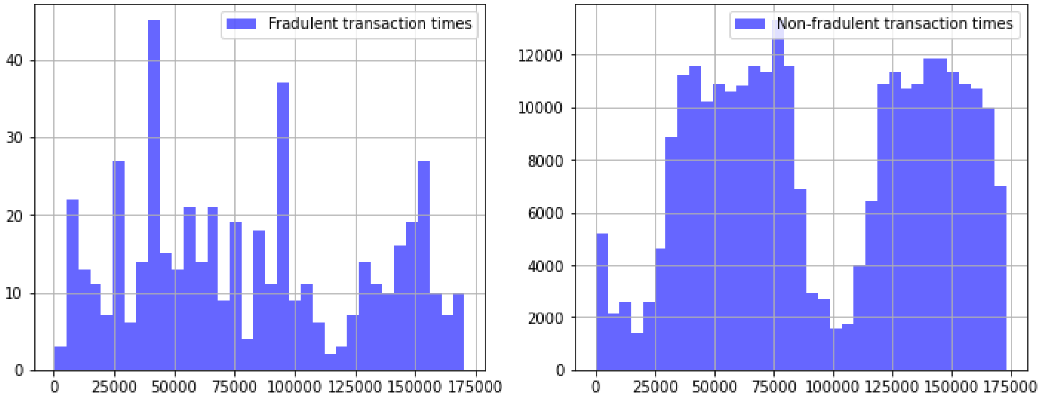

Figure 2: Transactions times of fraudulent and non-fraudulent transactions

The above can be cross verified by applying feature selection technique. For this study, variance threshold feature selection technique is used. Variance threshold is calculated based on probability density function of a particular distribution. If a feature has 95% or more variability, it is very close to zero and the feature may not help in the model prediction and those features are removed. Time was not selected in the outcome. The number of features from 28 to 23. The dataset is split into train and test where 80% is train dataset and 20% is test dataset.

Metric selection plays important part in validating the model. In this study, AUC (Area Under the Curve) is used. This is one of the good metrics to evaluate the score of classifiers since the calculation is based on the complete ROC (Receiver Operating Characteristic) curve and all possible classification threshold are implied. ROC-AUC validates that the method has good performance and identifies if transaction is risky or not.

Results:

In order to ensure there is no overfitting of data, cross validation needs to be performed. Cross validation ROC-AUC score is determined for different models and Random forest classifier score is 94.38. In random forest, each tree is trained by randomly sampling the subset of training dataset. Training is fast with large datasets and many features as each tree is trained independently of others. The model is evaluated using the test dataset and AUROC is 95.82

The results can be evaluated by confusion matrix. Confusion matrix shows the correct and incorrect prediction for each transaction type. First row, first column indicates the number of legitimate transactions that were predicted correctly. First row, second column indicates how many legitimate transactions were predicted as fraud. Similarly, second row indicates the fraudulent transactions that were erroneously predicted as legitimate. Hence, higher the diagonal values of confusion matrix, better would be the correct predictions.

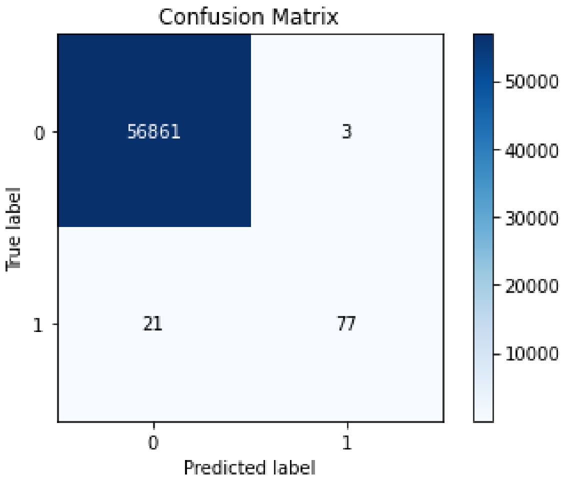

Figure 3: Classification matrix

Conclusion and future improvements:

Credit card fraud is major issue for financial institutions and they need to be quickly identified to minimize financial losses. The proposed random forest model AUROC metric of 95.82%. Log loss is another metric that is used and the score is 0.014. Low value of log loss indicates better performance of the models in future datasets. The number of false positives is less in this imbalanced dataset where the rate of fraud is less an 0.17%. Hence, this model can be used to predict fraudulent transactions. For future improvements, since the credit card fraud dataset is highly imbalanced, oversampling techniques can be used to increase the minority class. One such technique is SMOTE (Synthetic Minority Oversampling Technique). It can be applied to the training dataset and validate if that would improve the performance and thereby improve in predicting the fraudulent transactions.

Citations:

https://www.sciencedirect.com/science/article/pii/S2351978920314608

https://towardsdatascience.com/fraud-detection-the-problem-solutions-and-tools-dd8977b435c9