Credit Card Fraud Detection using Machine Learning

Nishchitha Manjuanth [gu7459]
Data Science & Business Analytics
Wayne State University, MI, USA

Dr. Dongxiao Zhu
Associate Professor, Dept of CSC
Wayne State University, MI, USA

Abstract **–The banking and financial systems are vulnerable to data attacks and fraudulent transactions. It is very important for the credit card companies to keep track on such activities and transactions. This project illustrates couple of machine learning models to recognize such fraudulent transactions. The goal here is to create a fraud detection model with high accuracy and minimal false rejections on transactions.**

I. Introduction

The illegal or unauthorized usage of one's credit card without the knowledge of the authorized cardholder is a fraud transaction, where the cardholder is charged for the items that haven't been purchased by him/her. This is a scenario which can be tackled using machine learning algorithms. Such fraudulent scenarios can be analyzed and could help protect from further occurrences of the same using this credit card fraud detection model. This model was trained with past credit card transactions which had a combination of normal and fraudulent transactions. This model would then be used to recognize a new transaction to be a fraud or a normal one. So, the aim of this project is to build a fraud detection model. The methodology which I have used are listed below.

II. Methodology

A. Data Preprocessing

The Credit Card Fraud Detection dataset consists 31 variables related to transaction details of a European cardholders out of which 28 variables are protected due to data security. The Time column show the duration gap between one transaction to the other, the Amount column shows the amount that was transacted, the Class column is a binary column where 1 represents a Fraudulent transaction and 0 represents a Normal transaction. This dataset has a total of 492 fraudulent transactions out of the 284,807 total transactions which makes the dataset an imbalanced one. The below plot shows the distribution of Fraud and Normal transactions.

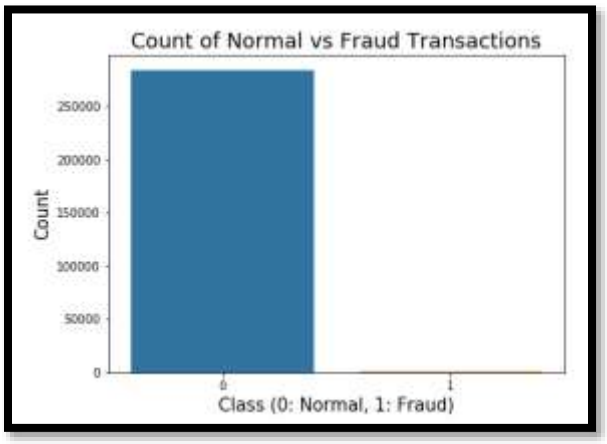

Fig1: Number of Transactions(Y-axis) to the type of Transaction(X-axis)

As there were no null values , no data imputation had to performed and the dataset was already transformed using PCA. The Time and Amount columns were scaled.

B. Data Resampling

As the dataset is highly imbalanced, data resampling must be performed. The following resampling methods were used-

Balanced Class Weight – This is a Sklearn package which proportionately weighs the class weights for unbalanced datasets.

GridSearchCV – This is used to perform hyper parameter tuning which estimates the optimized parameters by cross- validated grid search.

SMOTE – Synthetic Minority Oversampling technique is used to up sample the minority class. This approach increases the minority samples to get a better fit for the model.

Undersampling – This is a resampling technique which reduces the samples from majority class and matches it to the minority class.

C. Data Modeling

The steps which were performed for the model building are as below:

1. **Supervised Binary Classification**

For supervised learning, I have used standard scalar to scale the time and amount variables. The classification was done initially using Validation set, the best models with best performance were tested using the Test Dataset.

Logistic Regression Classifier: The given problem is an example of supervised binary classification as we need to classify the transactions to be either legitimate or fraudulent. So, Logistic regression is one of the classifiers that can be used for this problem.

RandomForest Classifier: This is an example for classification and regression. As our data is highly imbalanced, Random Forest algorithm is better resistant to overfitting and gives a good estimate of the error.

For both these supervised classifiers, resampling techniques as noted in the previous section were used to improve the performance of the model.

2. **Unsupervised Anomaly Detection**

For unsupervised learning, I have used PCA reduction on the time and amount columns. Also, I have performed two tailed z test for finding the most significant variables from the dataset.

Isolation Forest Algorithm: This is an unsupervised anomaly detection algorithm which returns a score of each anomaly sample.

The selection of feature is done arbitrarily and generates a minimum value and maximum value at random of the selected feature.

Local Outlier Factor: This is also an unsupervised anomaly detection algorithm that returns a score for each sample. The distance of nearest neighbors is used to estimate the sample data.

III. Results and Summary

A. Supervised Model: The table below shows the results. Random Forest was the best performing model followed by Logistic regression. I used various sampling techniques as noted before and the best results were obtained with down sampling. However, the main disadvantage of under sampling is that a lot of useful information is lost, thus this is not a practical approach. Amongst the oversampling techniques, the Balanced class with GridSearchCV was the best method.

Classifier Used	Methods Implemented	Accuracy	precision	Recall	F1Score	AUPRC
Logistic Regression	Oversampling using Balanced Class Weights	0.974	0.05	0.962	0.096	0.82
	Balanced Class + GridSearchCV	0.999	0.793	0.862	0.826	0.85
	SMOTE	0.972	0.047	0.962	0.09	0.83
	Pipeline(GridSearchCV +LR +SMOTE)	0.999	0.797	0.837	0.81	0.83
	Undersampling	0.944	0.969	0.922	0.945	0.98
Random Forest	Oversampling using Balanced Class Weights	0.999	0.953	0.762	0.847	0.88
	Balanced Class + GridSearchCV	0.999	0.955	0.813	0.878	0.89
	SMOTE	0.999	0.835	0.887	0.86	0.88
	Pipeline(GridSearchCV +RF +SMOTE)	0.999	0.91	0.887	0.898	0.88
	Undersampling	0.929	0.958	0.902	0.929	0.98

Table1: Classification Results Summary

B. Unsupervised Model: Isolation Forest is a better performer as compared to Local Outlier Factor. The below plot shows the count of normal and fraud transactions which were identified by both the classifiers.

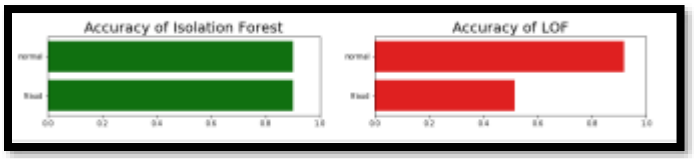

Fig2: Unsupervised Classification Results Summary

IV. Conclusion

The goal for any machine learning model is to find the best model for the given problem. Here in this work, I have shown a few different types of models. I feel, for Supervised Classification, Random Forest was the best model with Balanced + GridSearch resampling. For Unsupervised classification, Isolation Forest gave the best results.

V. References

[1]https://scikitlearn.org/stable/auto_examples/preprocessing/plot_all_scaling.html
[2] https://scikit-learn.org/stable/index.html
[3]https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173
[4] "Credit Card Fraud Detection Based on Transaction Behaviour -byJohn Richard D. Kho, Larry A.
[5] " A Comprehensive Survey of Data Mining-based FraudDetection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia
[6] "Survey Paper on Credit Card Fraud Detection by Suman" , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
[7]https://www.kaggle.com/mlgulb/creditcardfraud