Credit Card Fraud Detection using Machine Learning Algorithms

Madhuri Vemulapaty
Department of Electronics and Communication Engineering
PDPM Indian Institute of Information Technology Design and Manufacturing
Jabalpur,India
madhurivemulapaty@iiitdmj.ac.in

Dr.Irshad Ahmad Ansari
Department of Electronics and Communication Engineering
PDPM Indian Institute of Information Technology Design and Manufacturing
Jabalpur,India
irshad@iiitdmj.ac.in

***Abstract*— Credit card frauds are a source of huge losses to the banks every year. With the rise in the frauds due to increasing online transactions, fraud detection becomes an important problem. The present study uses historical labelled data to build models that classify input transactions as fraud or legitimate. A comparative study of three classifiers - logistic regression, support vector machine and random forest is performed on datasets with different under sampling ratios to choose the best performing classifier.**

Keywords*— *fraud detection, logistic regression, support vector machine, random forest, under sampling

I. Introduction

Billions of dollars are lost every year due to credit card fraud as reported in various articles [1-2]. With increase in online payments, there is also an increase in the number and variety of frauds as it is easier to hide location and identity on the Internet. This lead to a large amount of research on the fraud detection and prevention.

Machine learning techniques have been in use since 1990s for fraud detection with the algorithms becoming increasingly sophisticated today. Fraud detection methods are mainly classified into - misuse detection and anomaly detection [3]. Misuse detection uses historical data to label a transaction as fraud or legitimate. Both supervised and unsupervised methods can be used to achieve this. Anomaly detection uses transaction data of customers' behaviour to detect any anomaly and flag it as suspicious.

II. RELATED WORK

Research on using machine learning techniques for fraud detection has started in 1990's with Artificial Neural Network being one of the popular algorithms to be used. Ghosh et.al. [4] used a neural network trained on transaction data labelled fraud and not-fraud. A three layer feed-forward network was used to score transactions and then classify them based on a threshold value. Though the model performed well in terms of accuracy, it was limited by high computation time.

Srivastava et al. [5] used a Hidden Markov Model for anomaly detection. A HMM was initially trained with the cardholder's normal behaviour. If the HMM model does not accept the current transaction with sufficiently high probability, the transaction is flagged as fraudulent. However, only transaction amount is considered as a feature in this model.

Sahin et al. [6] used support vector machines and decision trees for fraud detection. The dataset is divided into three groups with different ratios of fraudulent and legitimate transactions. Seven decision trees and support vector machine based models are developed based on these datasets and the performance was evaluated. Results show that decision tree performs better than support vector machine. As the imbalance in dataset increases, accuracy is not a good performance metric and the model is limited by its use of accuracy as it performance metric. More recently, random forest algorithm [7] is used to solve the present problem. Here, random tree based random forest and CART based random forest were used to detect fraud. An initial comparison of the two was done and CART based random forest was chosen owing to its better performance. The technique of under sampling was used to tackle the problem of imbalanced data and choose the best ratio of under sampling. A bigger dataset is then used to demonstrate the effectiveness of the algorithm. Though the algorithm shows good results on small data, the effectiveness on highly imbalanced dataset still remains.

III. CLASSIFICATION ALGORITHMS

A. *Logistic Regression*

Logistic Regression is one of the most popular algorithms for classification problems owing to its simplicity and effectiveness [8]. It is a statistical model used to classify the output into one or more categories based on the relationship between the dependant and independent variables.

B. *Support Vector Machine*

A Support vector machine is a discriminative classifier that outputs an optimal hyperplane to categorise new inputs. These classifiers work in higher dimensional feature space which is a non-linear mapping of the input space making a non-linear classification task in the input space a linear classification in higher dimensions. A major advantage of SVM is its ability to work in higher dimensions without increasing computational complexity. SVM minimises overfitting by choosing a hyperplane with maximum margin of separation between the two classes. By choosing the appropriate kernel, cost parameter (C) and gamma values, the efficiency of SVMs can be greatly increased.

C. *Random Forest*

Random forest is an ensemble of decision tree classifiers. The output is the majority vote of the set of tree classifiers. The algorithm works well when the ensemble trees are dissimilar. This is achieved by building each tree using separate bootstrapped examples of data and using randomly selected subset of data attributes at each node while building individual trees. Random forests have been shown to perform better as compared to SVM and other techniques [9].

IV. PERFORMANCE METRICS

A. *AUC-ROC*

The Area under Curve Receiver Operating Characteristics is a performance metric for classification problems. It is a measure of the capability of the model to distinguish between two classes. Higher the score, better the performance. The receiver operating characteristics is plotted between True Positive Rate(TPR) versus False Positive Rate(FPR) where

TPR = True positives / (True positives + False negatives)

FPR = False positives / (True negatives + False positives)

This is considered to be an good metric to measure overall performance [10].

B. *F1 Score*

This gives the harmonic mean of precision and recall, where

Precision = True positives/ (True positives +False positives)

Recall = True positives/ (True positives + False negatives)

F1 = (2 * Precision * Recall)/ (Precision + Recall)

F1 score accounts for both precision and recall so maximizing the score maximizes both the metrics.

V. EXPERIMENT

The dataset for the current problem has been obtained from Kaggle [11]. The dataset is highly imbalanced with 0.172% being fraudulent cases and the rest legitimate. It contains only numerical input variables which are the result of a PCA transformation. Due to confidentiality issues, the original features and more background information about the data are not provided.

To deal with the problem of imbalanced data, the technique of random under sampling has proven to be effective [12]. Random under sampling is the random removal of samples from the majority class to balance the dataset. We first create a subsample of dataset with 1:1 under sampling ratio. We then find out the correlation between features to identify features with high positive and high negative correlation. We remove the extreme outliers of the dataset by removing samples that lie beyond 2.5 times the interquartile range.

The hyper parameters of support vector machine and random forest are tuned by Grid Search using AUC-ROC as a scoring metric. Grid Search builds a model for every combination of hyper parameters specified and evaluates each model. The hyper parameters of the model with highest score are then selected. A comparative analysis of the three classifiers is then made with the chosen performance metrics with K-Fold cross validation.

In order to model the real-world scenario where the percentage of fraudulent transactions is very small, we evaluate the performance of the three classifiers with different under sampling ratios - 15%, 10% and 5%.

VI. RESULTS AND DISCUSSION

The results of the three classifiers with different under sampling ratios is shown in the Table 1. As we can observe in Table 1, as the ratio of fraud cases decreases, the AUC-ROC and F1 score of logistic regression and support vector machine show a downward trend whereas those of random forest show better performance. Logistic regression and random forest show far better performance as compared to support vector machines with respect to both performance metrics.

The F-score of random forest is better than the other two with decreasing ratio of fraud cases. This is an important performance indicator as it shows that the system is correctly classifying frauds as well as minimizing errors in incorrect classification, both of which are extremely relevant to the real world scenario.

The performance of random forest can be further explored if the features are not anonymized and with further tuning of the hyper parameters. If the features of the data are known, feature extraction can be done to reduce the dimension of data. The tuning here is done by Grid Search whereas in further study, tuning can be improved by using Random Search followed by Grid Search to obtain better parameters.

TABLE I. PERFORMANCE OVER VARIOUS DATASETS

Classifier/ Under sampling ratios (Fraud: Legitimate)	**Logistic regression**		**Support Vector Machine**		**Random Forest**	
	AUC-ROC	**F1 score**	**AUC-ROC**	**F1 score**	**AUC-ROC**	**F1 score**
1:1	0.970	0.924	0.929	0.918	0.945	0.930
0.15:1	0.974	0.943	0.967	0.927	0.959	0.938
0.10:1	0.974	0.944	0.971	0.927	0.968	0.941
0.05:1	0.974	0.924	0.945	0.922	0.967	0.941

CONCLUSION

The study has compared the performance of the three classifiers – logistic regression, support vector machine and random forest for credit fraud detection. The technique of random under sampling has been used to deal with the problem of imbalanced data. To model the real-world scenario, performance of the algorithms on datasets with different under sampling ratios - 15%, 10% and 5%, was evaluated. Random forests show an overall better performance as compared to the other two. The performance of random forest can be further improved with better parameter tuning and feature extraction.

REFERENCES

[1] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, 235-249.

[2] Leonard, K. J. (1993). Detecting credit card fraud using expert systems. *Computers & industrial engineering*, *25*(1-4), 103-106.

[3] Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management*, *8*(2), 937-953.

[4] Ghosh, S., & Reilly, D. L. (1994, January). Credit card fraud detection with a neural-network. In *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on* (Vol. 3, pp. 621-630). IEEE.

[5] Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. *IEEE Transactions on dependable and secure computing*, *5*(1), 37-48.

[6] Şahin, Y. G., & Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines.

[7] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March). Random forest for credit card fraud detection. In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)* (pp. 1-6). IEEE.

[8] Shen, A., Tong, R., & Deng, Y. (2007, June). Application of classification models on credit card fraud detection. In *2007 International conference on service systems and service management* (pp. 1-4). IEEE.

[9] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602-613.

[10] Ling, C. X., Huang, J., & Zhang, H. (2003, August). AUC: a statistically consistent and more discriminating measure than accuracy. In *Ijcai* (Vol. 3, pp. 519-524).

[11] Machine Learning Group—ULB, Credit Card Fraud Detection (2018), Kaggle

[12] Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007, June). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning* (pp. 935-942). ACM.