Credit Card Fraud Detection Using Statistical & Machine Learning - A Comparative Study

Kamlesh Gupta
Department of Computing
Dublin City University
Dublin,Ireland
kamlesh.gupta3@mail.dcu.ie

Nivedita Mahato
Department of Computing
Dublin City University
Dublin,Ireland
nivedita.mahato2@mail.dcu.ie

Abstract - In this paper, we are focusing on credit card fraud detection in the real world. Generally, credit card fraud activities can happen in both online and offline. But in today's world, online fraud transaction activities are increasing day by day. So, to find the online fraud transactions, various methods have been used in the existing system. Here the credit card fraud detection is based on fraudulent transactions. In the proposed system, we compare Logistic Regression and Random Forest Algorithm for finding the fraudulent transactions and the accuracy of those transactions. Logistic Regression is a statistical model. In contrast, Random forest algorithm is based on supervised learning by using multiple decision trees for classification of the dataset. A confusion matrix is obtained for both models. The performance of Logistic Regression and the Random Forest Algorithm is evaluated based on the confusion matrix.

Keywords - Credit Card Fraud Detection, Transactions, Logistic Regression, Random Forest Algorithm.

I. INTRODUCTION

In today's world, most of the payments are often made using electronic devices. People use credit cards, and debit cards instead of cash for payments at the point of sale (POS) and can directly purchase products on shopping websites using their card data (E-commerce). However, the increasing use of digital transactions has led to new forms of crime, Hackers or fraudsters attempts to exploit the card data of legitimate users to make payments on their behalf. In spite of the number of authorization techniques in place, credit card fraud cases have not yet reduced. Also, as the locations are hidden, Fraudsters prefer to use the Internet platform for this, and the increase in credit card fraud has a massively impacted the financial industry [3]. According to Nilsson Report, the Global losses of credit card fraud reached $16.31 billion in 2014, and it is estimated to exceed $35 billion in 2020 [4].

To identify suspicious credit card transactions, payment gateways use serval of identification techniques to discard the fraud-initiated transaction. Among them, two mechanisms most used are fraud prevention and fraud detection. Fraud prevention is a proactive approach to preventing fraud from happening in the first place. On the other hand, fraud detection is needed when a suspicious transaction has already taken place in order to discard the transaction or to penalize the fraudster.

As machine learning and data mining techniques are widely utilized to deal with cyber-criminal cases, researchers often use those models to determine credit card fraud activities as well. In many literature Machine learning techniques like support vector machine, Naive Bayes, Artificial neural network, decision tree has been used in credit card fraud detection. This paper seeks to perform statistical analysis on the credit card dataset and utilize logistic regression and Random forest technique and evaluate their performance on credit card fraud data.

The rest of this paper is arranged as follows: Section II gives a detailed review of the past literature on credit card fraud detection techniques and performance comparison. Section III gives a brief description of the dataset and the exploratory analysis performed on it. Section IV provides the details description of our hypothesis. Section V describes the methodological contributions of the paper and Section VI concludes the comparative study and suggests future areas of research [13].

II. RELATED WORKS

In recent studies, the data mining technique is one notable method for credit fraud detection. Billions of dollars are lost due to credit card fraud every year [3]. Identification of those transactions which are fraudulent into two classes of legitimate (genuine) and fraudulent transactions is the process of credit card detection [1]. Credit card fraud detection is based on an analysis of a card's spending behavior. Many techniques have been applied to credit card fraud detection, naïve Bayes, k-nearest neighbor and logistic regression classifiers [1], A feedback mechanism was built to make use of the True label information from transactions and form a fraud detection method. Behavioral patterns from similar cardholders were created as a dataset for the experiment. [2], support vector machine, Naive Bayes [3], the supervised based classification using Bayesian network classifiers namely K2, Tree Augmented Naïve Bayes (TAN), and Naïve Bayes, logistics and J48 classifiers [4]. After pre-processing the dataset using normalization and Principal Component Analysis, all the classifiers achieved more than 95.0% accuracy compared to results attained before pre-processing the dataset [4]. In [5] Nurul et al. have a comparison between various machine learning algorithm and the study suggest that stacking classifier, which uses Logistic Regression as meta classifier is most promising for predicting. Deep transfer learning approaches for credit card fraud detection and focuses on transferring classification models [6]. Suresh et al. experimented using Random Forest

Algorithm (RFA) for finding the fraudulent transactions and the accuracy of those transactions and gained 90% accuracy [7]. Pavan et al. did a study of multiple machine learning algorithms such as Neural Network (NN), rule induction techniques, fuzzy system, call trees, Support Vector Machines (SVM) [4], Logistic Regression, Local Outlier Factor (LOF), Isolation Forest, K-Nearest Neighbor, Genetic algorithms for fraud detection [4]. Many attempts have been made to improve credit card fraud detection but lack somehow due to several problems as fraudulent behavior profile are dynamic, that is fraudulent transactions tend to look like legitimate ones, credit card transaction datasets are rarely available and highly imbalanced (or skewed), optimal feature selection for the models, suitable metric to evaluate performance of techniques on skewed credit card fraud data[1]. Credit card fraud detection algorithms performance is greatly affected by the type of sampling approach used, selection of variables and detection techniques used.

III. DATASET AND EXPLORATORY ANALYSIS

The dataset used in this project contains the transaction details made by the European cardholders in September 2013. This dataset describes transactions that took place in two days, where out of 284,807 transactions, we have 492 transactions marked as frauds, and it makes up 0.172% of total transactions. This dataset is highly imbalanced and positively skewed [14]. It only contains numerical input variables resulting from the Principal Component Analysis (PCA) feature selection transformation resulting in 28 principal components. So, in total, 30 input features have been utilized as a part of this study. The detailed background of each feature has been removed as part of the data cleaning process due to confidentiality issues. The time attribute includes the seconds for each transaction and the first transaction in the dataset. The amount attribute is the amount transaction, and it can be used for cost-sensitive learning, for instance. The attribute 'Class' is the output class for the binary classification, and it takes value 0 and 1 as the value to indicate fraud and non-fraud [14]. Also, in this feature the fraudulent transactions seem to have higher mean value than non-fraudulent ones meaning that this feature would likely be useful to use in the predictive model. However, the median for this is higher for the legitimate ones, which implies that the distribution of values for class "0" is left-skewed [10].

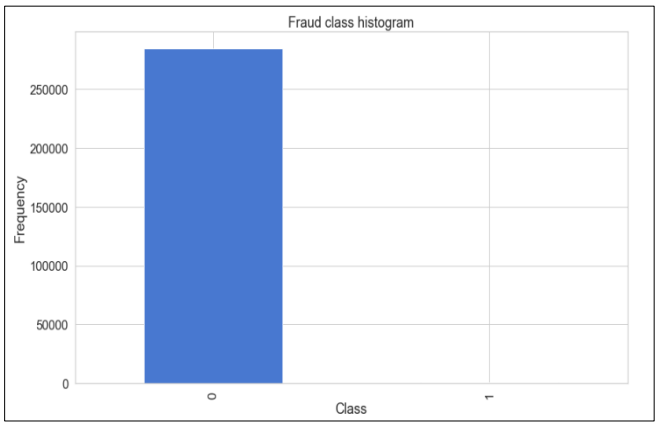

Fig 1. Fraud Class Distribution

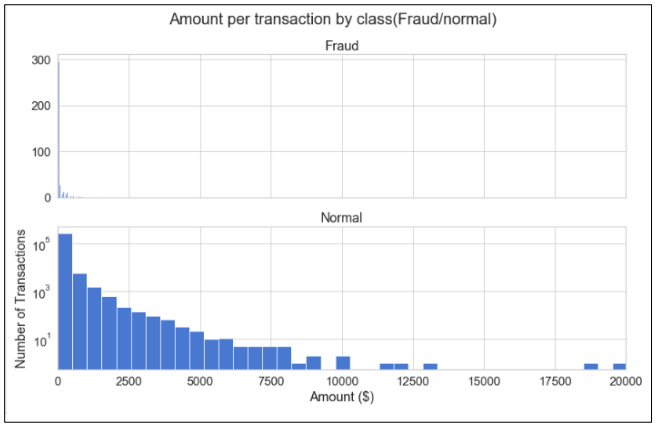

Fig 2. Fraud vs Normal Transaction

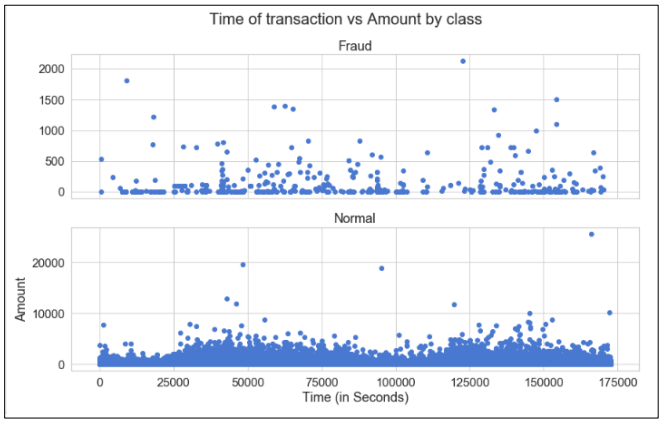

Fig 3. Time of the Transaction vs Amount by Class.

IV. COMPARATIVE ANALYSIS

From the above previous work, it has shown us that various methods and techniques have been studied for detecting fraudulent transaction in a credit card. Hence, we want to propose our hypothesis as below:

Ho: Machine learning gives a better result than the statistical model in fraud detection.
Ha: Machine learning does not give a better result than the statistical model in fraud detection.

For analyzing our null hypothesis, we are going to start with a simple statistical model and move on to more complex ones so that we have a rough idea of which model performs better with our data. Also, it is important to consider the trade-off between model accuracy and model complexity. For example, it might be the case that having a 10-layer neural network which trains for two days on a GPU cluster and is 90% accurate will not be preferred to a simple model with short inference times which achieves an accuracy of 85% which is enough for a given task. We are going to perform our analysis using two methods.

V. EXPERIMENT

Method 1 - Logistic regression is a simple regression model whose output is a score between 0 and 1. It is achieved by using the logistic function [9]

$$g(z)=11+\exp(-z)g(z)=11+\exp(-z)$$

Where: $z = \beta Txz = \beta Tx$

The model can be fitted using gradient descent on the vector beta. We used 70% of the data is used for training and 30% used for the testing set. Data was balanced by using an under-sampling technique with five iterations; steady values of parameters are achieved. Gradient ascent incrementally updates the classifier as new data enter in then all at once. It starts with all weights set to 1. Then for every feature value in the dataset, the gradient ascent is calculated. The gradient ascent is used in this study because, given the large size of data, it updates the weights using only one instance at a time, thus reducing computational complexity.

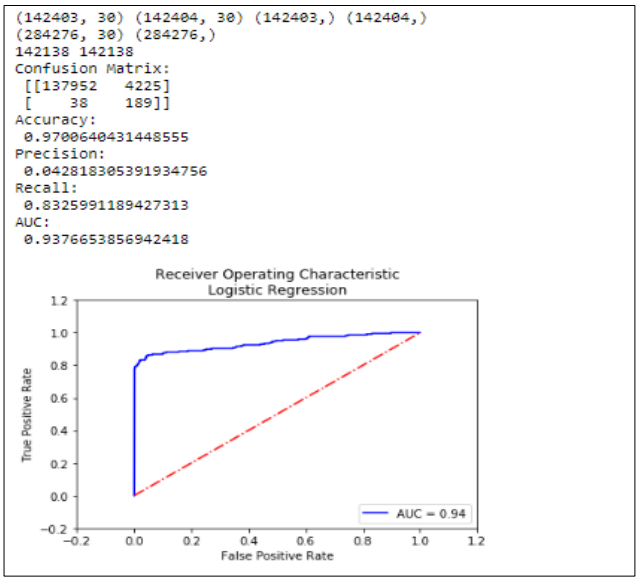

Fig. 4 Logistic Regression.

We received 97% accuracy using logistic regression with Recall – 83.25%, Precision – 0.04 % and AUC – 93.76%.

Method 2: The random forest algorithm is one of the supervised learning models; it uses labelled data to learn how to classify unlabeled data. Regression and classification problems can be solved by using the Random Forest Algorithm, making it a diverse model that is widely used by various researchers.

The random forest makes hundreds or thousands of decision trees. It trains each one on slightly different observations set. After splitting nodes in each tree, it considers a limited number of the features. Averaging the predictions of each tree will be done for the final predictions. We have used entropy for branching the nodes in a decision tree. It uses the probability of an outcome to decide on how the node should branch.

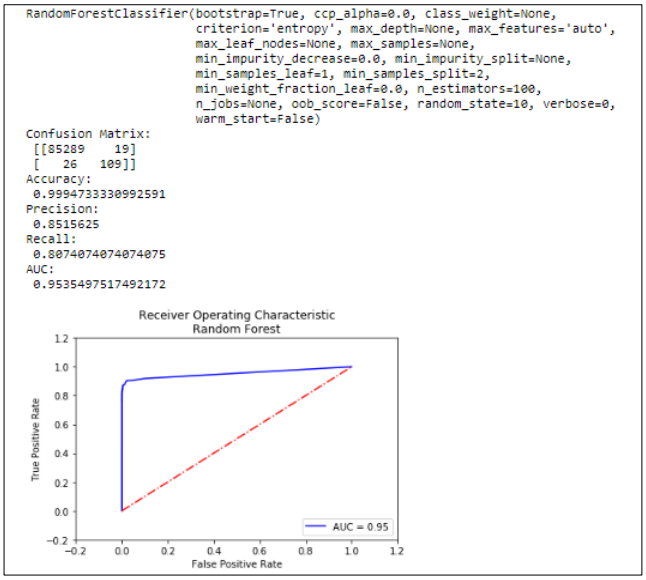

Fig. 5 Random Forest.

We received 99.94% accuracy using the Random Forest Algorithm with Recall – 80.74%, Precision – 85.15 % and AUC – 95.35%.

VI. CONCLUSIONS

In this paper, we have performed a statistical analysis on the credit card dataset and utilize logistic regression and Random forest technique to detect the fraud transaction. Furthermore, we examined their comparative performance on credit card fraud dataset. The main reason for evaluating these two techniques is because very less comparison has been made in past literature.

The contribution of the paper is summarized in the following:

1. Performed an exploratory analysis to understand the dataset for fraud & non-fraud and validate if it contains any NULL values.

2. Performed over-sampling on the imbalanced dataset

3. Created trained and test data.

4. The performance of the two approaches is examined using attributes such as accuracy, precision and AUC.

We observed that even a straightforward logistic regression model could achieve good recall, while a much more complex Random Forest model improves upon logistic regression in terms of AUC. Hence this study shows that we fail to reject our null hypothesis

Future work can be carried out using balanced historical complex data and comparing the performance of Logistic regression and Random forest precision and accuracy with other machine learning methods.

References

[1] Dighe, D., Patil, S. and Kokate, S., 2018, August. Detection of Credit Card Fraud Transactions Using Machine Learning Algorithms and Neural Networks: A Comparative Study. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). IEEE. Cited by 1.

[2] Jiang, C., Song, J., Liu, G., Zheng, L. and Luan, W., 2018. Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism. IEEE Internet of Things Journal, 5(5), pp.3637-3647. Cited by 13.

[3] Randhawa, K., Loo, C.K., Seera, M., Lim, C.P. and Nandi, A.K., 2018. Credit card fraud detection using AdaBoost and majority voting. IEEE access, 6, pp.14277-14284. Cited by 39.

[4] Yee, O.S., Sagadevan, S. and Malim, N.H.A.H., 2018. Credit card fraud detection using machine learning as data mining technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4), pp.23-27. Cited by 14.

[5] Dhankhad, S., Mohammed, E. and Far, B., 2018, July. Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 122-125). IEEE. Cited by 14.

[6] Lebichot, B., Le Borgne, Y.A., He-Guelton, L., Oblé, F. and Bontempi, G., 2019, April. Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection. In INNS Big Data and Deep Learning conference (pp. 78-88). Springer, Cham. IEEE. Cited by 4.

[7] Kumar, M.S., Soundarya, V., Kavitha, S., Keerthika, E.S. and Aswini, E., 2019, February. Credit Card Fraud Detection Using Random Forest Algorithm. In 2019 3rd International Conference on Computing and Communications Technologies (ICCCT) (pp. 149-153). IEEE.

[8] Kumar, P. and Iqbal, F., 2019, April. Credit Card Fraud Identification Using Machine Learning Approaches. In 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) (pp. 1-4). IEEE.

[9] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. and Bontempi, G., 2017. Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), pp.3784-3797. Cited by 44.

[10] Porwal, U. and Mukund, S., 2018. Credit Card Fraud Detection in e-Commerce: An Outlier Detection Approach. arXiv preprint arXiv:1811.02196. Cited by 5.

[11] Lebichot, B., Le Borgne, Y.A., He-Guelton, L., Oblé, F. and Bontempi, G., 2019, April. Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection. In INNS Big Data and Deep Learning conference (pp. 78-88). Springer, Cham. Cited by 5.

[12] Bhatla, T.P., Prabhu, V. and Dua, A., 2003. Understanding credit card frauds. Cards business review, 1(6). Cited by 105.

[13] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M. and Anderla, A., 2019, March. Credit Card Fraud Detection-Machine Learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.

[14] Campus, K., 2018. Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models. International Journal of Pure and Applied Mathematics, 118(20), pp.825-838.