International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 05 - Issue 06 || June 2020 || PP. 74-80

Credit card fraud detection through model testing using Orange Tool

Fakhra Akhtar
Faculty of Computer Science an Information Technology, RIPHA University Lahore, Pakistan
2nd-June-2020

Muhammad Tauseef Hanif
COMSATS University Islamabad, Lahore-Campus

Faizan Ahmed Khan
COMSATS University Islamabad, Lahore-Campus

Abstract: Due to the tremendous growth of technology, digitalization has become the key angle in the financial division. As online exchange builds, the extortion rate develops at the same time. MasterCard extortion is an exhaustive term for misrepresentation perpetrated utilizing an installment card, for example, a MasterCard or charge credit can happen, where the client himself forms an installment to another record which is constrained by a lawbreaker, or unapproved, individual or where the record holder doesn't give approval to the installment to continue and the exchange is done by an outsider. Frauds caused by Credit Cards have costs consumers and banks billions of dollars globally. Even after numerous mechanisms to stop fraud, fraudsters are continuously trying to find new ways and tricks to commit fraud.
The dataset used in this paper relates to credit card fraud detection. The algorithms used in this paper are linear regression, Random Forest Algorithm and Ad boost The data set is taken from UCI ML repository. The objective of this paper is to develop a unique credit card fraud detection system by using a comparison of various target variable algorithms, and by comparing various models, the target variable is denoted by V1. There are thirty-one attributes in the dataset. Evaluation results are determined using orange tool.
The results showed that ad boost model holds highest value i.e. 0.984 followed by random forest i.e. 0.975 whereas linear regression holds lowest value which is 0.396. probability difference is negligible among models.
Keywords: Credit card fraud detection, Orange, Ad boost

INTRODUCTION

Credit card fraud is when someone uses your credit card to make a purchase you didn't authorize. This activity can happen in different ways like if you lose your credit card or have it stolen, it can be used to make purchases or other transactions, either in person or online. It is a problem that has affected the entire consumer credit industry and is one of the fastest-growing types of fraud and is most difficult to prevent. It involves the duplication of credit cards. Criminals have been able to use technology, with relative ease, to produce fraudulent versions of existing credit cards. The Internet helped this scheme to grow. Some criminals sell the magnetic strips found on many cards or the technology to duplicate the information from a valid credit card. These magnetic strips contain all the information a fraudster needs: names, account numbers, credit limits, plus other identifying information. Using a computer system and the right equipment, a fraudster can create a fraudulent credit card with ease. Fraudsters also use technology to create fictitious cards, which are more advantageous because there is no person truly responsible for the account. The credit card companies will notice that the account is not being paid and they will attempt to contact the account holder, but no one exists.

The first phase of the paper explains about literature review regarding to the defined dataset i.e. credit card fraud detection The second phase deals with the methodology data analysis tool and classification of algorithms. The third part discusses the research results based on the classification algorithms used in the Orange Tool. Fourth phase deals with Discussion, conclusion and references.

LITERATURE REVIEW

Kuldeep Randhawa et al. [1] proposed a technique using machine learning to detect credit card fraud detection. Initially, standard models were used after that hybrid models came into picture which made use of AdaBoost and majority voting methods. Publically available data set had been used to evaluate the model efficiency and another data set used from the financial institution and analyzed the fraud. Then the noise was added to the data sample through which the robustness of the algorithms could be measured. The experiments

were conducted on the basis of the theoretical results which show that the majority of voting methods achieve good accuracy rates in order to detect the fraud in the credit cards. Thus, it was concluded that the voting method showed much stable performance in the presence of noise. Abhimanyu Roy et al. [2] proposed deep learning topologies for the detection of fraud in online money transaction. They have used high performance distributed cloud computing environment. The study proposed by the researchers provides an effective guide to the sensitivity analysis of the proposed parameters as per the performance of the fraud detection. The researchers also proposed a framework for the parameter tuning of Deep Learning topologies for the detection of fraud. This enables the financial institution to decrease the losses by avoiding fraudulent activities. Shiyang Xuan et al. [3] used two types of random forests which train the behavior features of normal and abnormal transactions. The researcher compares these two random forests which are differentiated on the basis of their classifiers, performance on the detection of credit card fraud. The data used is of an e-commerce company of China which is utilized to analyze the performance of these two types of random forests model. In this paper, the author has used B2C dataset for the identification and detection of fraud from the credit cards. Therefore, the researcher concluded from the result that the proposed random forests provide good results on small dataset but there are still some problems like imbalanced data which makes it less effective than any other dataset. Sharmistha Dutta et al. [4] presented a study on the commonly found crime within the credit card applications. There are certain issues faced when the existing non-data mining approaches are applied to avoid identity theft. A novel data mining layer of defense is proposed for solving these issues. For detecting the frauds within various applications, two algorithms named Communal Detection and Spike Detection which generate novel layer. There is a large moving window, higher numbers of attributes and numbers of link types available which can be searched by CD and SD algorithms. Thus, results can be generated by the system by consuming a huge amount of time. Therefore, it is not possible to properly demonstrate the concept of adaptability. These issues can be resolved by making certain enhancements in the proposed algorithm in future work. (5) FRANCISCA (2011) in his research consider Data mining being popularly used to combat frauds because of its effectiveness. It is a well-defined procedure that takes data as input and produces models or patterns as output. Neural network, a data mining technique was used in his study. The design of the neural network (NN) architecture for the credit card detection system was based on unsupervised method, which was applied to the transactions data to generate four clusters of low, high, risky and high-risk clusters. The self-organizing map neural network (SOMNN) technique was used for solving the problem of carrying out optimal classification of each transaction into its associated group, since a prior output is unknown. The receiver-operating curve (ROC) for credit card fraud (CCF) detection watch detected over 95% of fraud cases without causing false alarms unlike other statistical models and the two-stage clusters. The results show that the performance of CCF detection watch is in agreement with other detection software, but performs better. The results of the study show the fact that The CCF detection system was designed to run at the background of existing banking software and attempt to discover illegitimate transactions entering on real-time. basis. This proved to be very effective and efficient method of discovering fraudulent transactions.

METHODOLOGY

Artificial intelligence is considered as a use of computerized reasoning (AI) that gives frameworks the capacity to naturally take in and improve as a matter of fact without being customized. It centers around the advancement of PC programs that can get to information and use it for themselves in this research linear regression, random forest and Ad boost are used.

I. DEFINATIONS OF ALOGRITHMS

Linear Regression.

Linear regression quantifies the relationship between one or more predictor variables and one outcome variable. It is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as multiple regression, multivariate regression, ordinary least squares (OLS), and regression.

Random Forest

The Random forest are considered as a learning strategy for characterization, relapse and different assignments that work by building a mass of choice trees at preparing time. The Random Forest Classifier is characterized as set of choice trees from haphazardly chosen subset of preparing set (5) It alternatively creates two extra snippets of data: a proportion of the significance of the indicator factors, and a proportion of the inner structure of the information.

Adaboost

AdaBoost, short for "Adaptive Boosting", is the first practical boosting algorithm proposed by Freund and Schapire in 1996. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one.

The Data

Data used in this research paper relates to credit card fraud detection This data will be helpful for the researcher to identify various frauds caused by credit cards in order to test the various models namely linear regression, random forest and adaboost Orange tool was used to produce the result.

Orange

Orange is an open source data visualization and data analysis tool for data mining through visual programming or Python scripting. The tool has components for almost all well-known machine learning algorithms, add-ons for bioinformatics and text mining as well as features for data analytics also. So, for researchers it is a one stop solution for pre-processing of dataset, visualization of dataset using graphs, all inbuilt machine learning algorithms, test and score feature for measuring accuracy of algorithm on different datasets along with many more fantastic features. This program provides a platform for experiment selection, recommendation systems, and predictive modeling and is used in biomedicine, bioinformatics, genomic research, and teaching. In data science, it is used as a platform for testing new machine learning algorithms and for implementing new techniques in genetics and bioinformatics.

RESULTS

In order to test the model for the defined dataset orange tool was used the results are shown in the form of various tables and graphs.

II. Table 1 Model evaluation result

Model	MSE	RMSE	MAE	R2
Random forest	**0.095**	**0.308**	**0.137**	**0.975**
Linear regression	**2.317**	**1.522**	**0.890**	**0.396**
AdaBoost	**0.063**	**0.250**	**0.110**	**0.984**

Table 1 shows the evaluation results of various models it is observed that adaboost model holds highest value i.e. 0.984 followed by random forest i.e. 0.975 whereas linear regression holds lowest value which is 0.396

a. Table 2 Training and simulation errors

	Random forest	Linear regression	AdaBoost
Random forest		**0.000**	**1.000**
Linear regression	**1.000**		**1.000**
AdaBoost	**0.000**	**0.000**	

In table 2 the probabitlies describe the fact that the score for the model in the row is higher than that of the model in the other two column, small numbers describe the fact that the probability difference is negligible.

B Table 3 prediction values of the models against the target variable

V1	Linear Regression	Random Forest	AdaBoost	
-1.35981	-0.885905	-0.831667	-1.04726	1
1.22966	0.167993	1.2283	1.23418	1
1.25	0.15956	1.25345	1.25308	1
1.10322	0.227526	1.10461	1.10336	1
-1.94653	-1.89141	-1.59519	-1.74191	1
1.17328	0.143433	1.16999	1.16784	1
-0.529912	-0.604701	-0.520804	-0.529912	1

Following table shows prediction values of the models against the target variable V1 linear regression shows highest value i.e. 88 percent whereas random forest has 83 percent the lowest value is of adaboost which is 52 percent

Figure 1 shows the graphical representation of line chart and its results

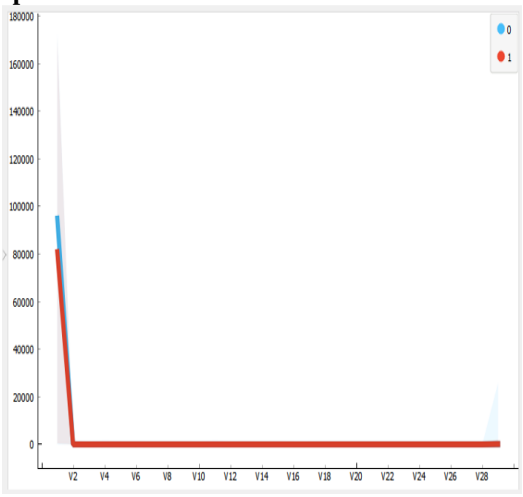
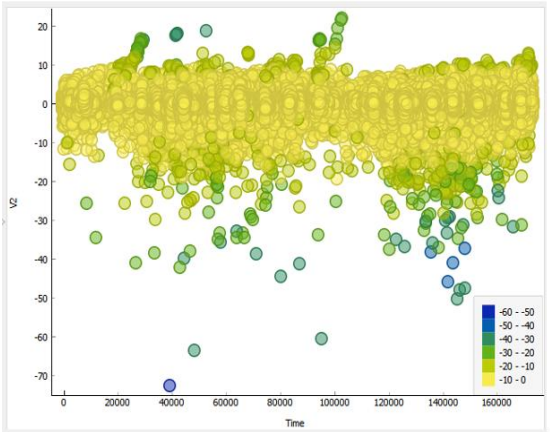


Figure 2 shows the graphical representation of scatter plot and its results



At x-axis time is taken whereas on y-axis v2 variable is plotted It is observed that mostly results are in the range of 0-10 which means there is high predication against time

Figure 3 shows graphical representation of random forest model against V1 variable the highest frequency shown is 80000

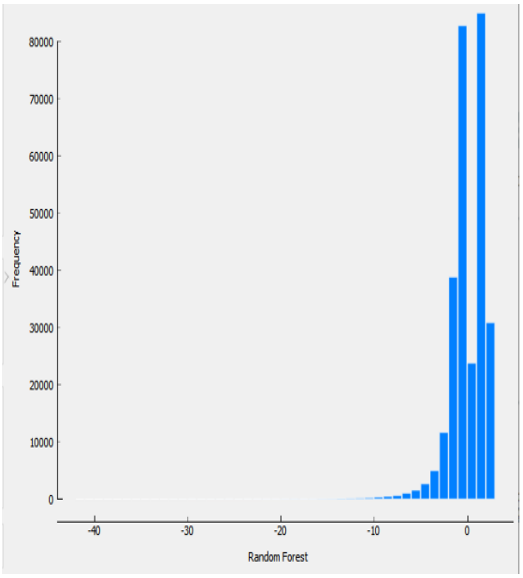


Figure 4 shows the graph of linear regression which is used to measure the highest frequency value of the defined data set against the target variable i.e. V1 it is observed that highest frequency value is 160000

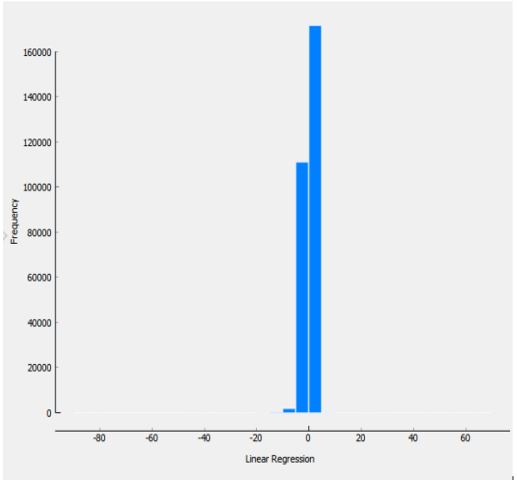
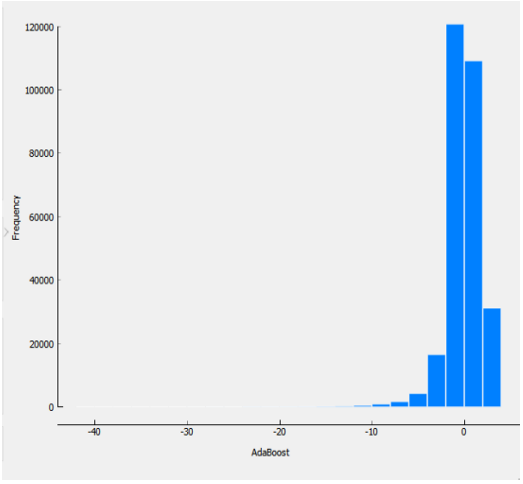


Figure 5 shows the graph of ad boost model set against the target variable V1 and frequencies it is observed that highest frequency value is 120000



DISCUSSION

Based on assumption it is observed that ad boost model holds highest value i.e. 0.984 followed by random forest i.e. 0.975 whereas linear regression holds lowest value which is 0.396. probability difference is negligible among models. Similarly, prediction values of the models against the target variable V1 linear regression shows highest value i.e. 88 percent whereas random forest has 83 percent the lowest value is of adaboost which is 52 percent in scatter plot It is observed that mostly results are in the range of 0-10 which means there is high predication against time.

CONCLUSION AND RECOMMENDATION

The results show the fact that linear regression and random forest are the best model to predict the fraud against the credit cards The results showed that ad boost model holds highest value i.e. 0.984 followed by random forest i.e. 0.975 whereas linear regression holds lowest value which is 0.396. probability difference is negligible among models. This research recommends that Credit card fraud and debit card fraud are huge threats to merchants. There should be a proper mechanism to prevent the frauds by improving security protocols and protection of cardholder's identity by introducing unique security pin cords and OTP these frauds can be prevented with the right approaches, technologies, tools, and awareness.

REFERENCES

[1] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim and Asoke K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," IEEE Access, vol. 6, pp. 14277-14284, 2018.

[2] A. Roy and J. Sun and R. Mahoney and L. Alonzi and S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," in Systems and Information Engineering Design Symposium (SIEDS), pp. 129-134, 2018.

[3] Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang and Changjun Jiang Shiyang Xuan, "Random Forest for Credit Card Fraud Detection," in IEEE 15th International Conference On Networking, Sensing and Control (ICNSC), pp.1-6, 2018.

[4] S. Dutta, A. K. Gupta and N. Narayan, "Identity Crime Detection Using Data Mining, "3rd International Conference on Computational Intelligence and Networks (CINE), Odisha, pp. 1-5, 2017.

[5] Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. R news,

[6] 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.

[7] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

[8] Aleskerov, E., Freisleben, B., & Rao, B. (1997, March). Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr) (pp. 220-226). IEEE.

[9] Stolfo, S., Fan, D. W., Lee, W., Prodromidis, A., & Chan, P. (1997, July). Credit card fraud detection using meta-learning: Issues and initial results. In AAAI-97 Workshop on Fraud Detection and Risk Management (pp. 83-90).

[10] Raj, S. B. E., & Portia, A. A. (2011, March). Analysis on credit card fraud detection methods. In 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET) (pp. 152-156). IEEE.

[11] Bhatla, T. P., Prabhu, V., & Dua, A. (2003). Understanding credit card frauds. *Cards business review*, *1*(6), 1-15.

[12] Ogwueleka, F. N. (2011). Data mining application in credit card fraud detection system. *Journal of Engineering Science and Technology*, *6*(3), 311-322.

[13] Chen, R. C., Chen, T. S., & Lin, C. C. (2006). A new binary support vector system for increasing detection rate of credit card fraud. *International Journal of Pattern Recognition and Artificial Intelligence*, *20*(02), 227-239.

[14] Alfuraih, S. I., Sui, N. T., & McLeod, D. (2002). Using trusted email to prevent credit card frauds in multimedia products. *World Wide Web*, *5*(3), 245-256.

[15] Sethi, Neha, and Anju Gera. "A revived survey of various credit card fraud detection techniques." *International Journal of Computer Science and Mobile Computing* 3, no. 4 (2014): 780-791.

[16] Patel, R. D., & Singh, D. K. (2013). Credit card fraud detection & prevention of fraud using genetic algorithm. *International Journal of Soft Computing and Engineering*, *2*(6), 292-294.

[17] Wang, D., Chen, B., & Chen, J. (2019). Credit card fraud detection strategies with consumer incentives. *Omega*, *88*, 179-195.

[18] 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.

[19] Raj, S. B. E., & Portia, A. A. (2011, March). Analysis on credit card fraud detection methods. In *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)* (pp. 152-156). IEEE.

[20] Li, Y., & Zhang, X. (2005). Securing credit card transactions with one-time payment scheme. *Electronic Commerce Research and Applications*, *4*(4), 413-426.

[21] Singh, A., & Jain, A. (2019). Adaptive credit card fraud detection techniques based on feature selection method. In *Advances in computer communication and computational sciences* (pp. 167-178). Springer, Singapore.

ACKNOWLEDGMENT

I would like to thank Prof. Dr Imran Ahmad from the department of Computer Science and Information Technology, RIPHA University Lahore, Pakistan.