Comparing Different Models for Credit Card Fraud Detection

Bhupendra Singh[1], Mehul Mahrishi[2]
[1]Software Development, HCL Technologies Ltd, Noida, 201301(INDIA)
[2]Department ofInformation Technology, Swami Keshvanand Institute of Technology, Management & Gramothan Jaipur-302017 (INDIA)
Email: bbhupendra007@gmail.com, mehul@skit.ac.in

Received 18.07.2019 received in revised form 27.07.2019, accepted 06.10.2020

Abstract- This paper incorporates the Credit Card Fraud Detection models to study and identify legitimate and fraud transactions. This research intends to recognize the false transactions while avoiding incorrect fraud classifications. The informational collection or dataset (Credit Card Fraud Detection) utilized in the proposed work is given by Kaggle which can be at https://www.kaggle.com/mlg-ulb/creditcardfraud. Before uploading on the website (Kaggle.com), these features are renamed and re-defined as PCA (Principal Component Analysis). There are general features in which 28 out of them are renamed as V1 through V28 (all numeric qualities). Rest three of the features showcase the time, calculated amount and whether that transaction was fraudulent or not. The response variable is 1 for a false transaction and 0 for a safe transaction. The chosen data set does not contain missing qualities. The dataset contains 284,807 transactions in which most of the transactions are very small and very few of the transactions come even closer to the maximum.

Different algorithms are implemented in this study. Python Machine Learning libraries are used to perform those algorithms. The model studied in this research work are K-Nearest Neighbour, logistic regression, random forest model, XGBoost model. As the XGBoost is showing more accuracy than other models. Out of these algorithms, XGBoost model is preferable over the Random Forest model and Logistic Regression model.

Keywords –Credit Card Fraud, Fraud Detection Comparisons, XGBoost, Random Forest, Machine Learning.

1. INTRODUCTION

Credit Card Fraud Detection is quick moving among individuals. Nowadays, it is a serious threat to online or offline buyers who uses their cards to avail of different kind of services. There can be 2 types of misrepresentation of Credit Card – Offline and Online.

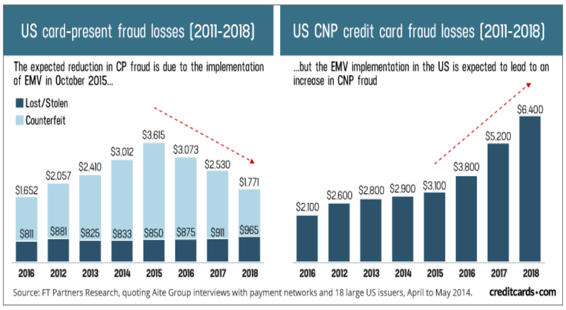
Offline fraud is put together with the use of a stolen card at client confronting venue or by the call center. Overall, the Card Issuing Authority can jolt it before being used in a bogus manner. Online misrepresentation is submitted through online transactions, e-shopping or cardholder not present. In 2018, over 10% of financial companies such as banks, credit-card authorities, etc. have faced the problem of data breaches. The loss of individual data straightforwardly adds to developing misrepresentation misfortunes for banks and merchants. It is important to collect key details for reasons of breaches to avoid losses from the financial service organization. This information is planned to help extortion supervisory groups figure out where holes exist in the security issues in this industry.

In recent years, fraud is top of mind with many people. Lawbreakers are progressively utilizing tricks to fool individuals into uncovering their own subtleties or parting with their money. Raising open mindfulness is vital to beating the fraudsters. In 2015, Money related extortion had misfortune crosswise over cards, banking, etc. which was totaled £755 million. It was a gain of 26 percent as compared to 2014. Now a days, a huge percentage of card fraud are registered from Card Not Present (CNP) payments i.e. payment through websites or online payments, which is 73%, whereas 19% of the fraud transactions are registered from Point of Sale (POS) terminals and a small fraction of 8% was recorded from Automated Teller Machines (ATMs). In 2018, Card Not Present (CNP) frauds are emerging because of the non-presence of the customer in online transactions. In the US, there was the highest sale on eCommerce businesses as 77% of traders are selling products online but the percentage of CNP fraud also increases as shown in the figure.

2. OBJECTIVE

With the developing use of credit card transactions, financial frauds have likewise been expanded prompting the loss of peoples. Having a proficient fraud detection identification strategy has turned into a need for all banks to limit such misfortunes. Credit Card fraud detection system is challenging, because the dataset provided for fraud detection is very unbalanced, as the quantity of false exchanges is a lot littler than the real ones. Thus, many of fraud detection models got failed due to these data sets. This aim of this study is to enhance the performance of the minority of credit card fraud on the dataset available. So, K-means clustering, logistic regression, random forest, and XGBoost models are performed. The objectives of the study are:

- To recognize the various kinds of charge card frauds.
- To study the previously implemented techniques that has-been utilized in fraud-detection.
- To compare and analyze recently published at and investigate as of findings in credit card fraud detection.



Source: FT Partners Research, quoting Aite Group interviews with payment networks and 18 large US issuers, April to May 2014.

Figure 1: Credit Card Fraud Losses

3. PROBLEM FORMULATION

To address this problematic situation, two main factors have to be considered. Firstly, we would take the random under-sampling technique into the picture to generate the training dataset along with balanced class distribution which will result into the detection of illicit transactions by forcing the algorithm so as to achieve high performance.

As far as the performance is concerned here, we are not supposed to depend on accuracy. While we will mark the performance through the best purpose of Receiver Operating Characteristics - Area under the Curve abbreviated as ROC-AUC Performance Measure. Most importantly, the ROC-AUC scores a value between 0 and 1, where 1 denotes to perfect performance score and 0 is the worst. If ROC-AUC score falls more than 0.5, then it is a sign of achieving higher performance than random guessing.

To generate our balanced training dataset, same quantity of fraud and non-fraud transactions are chosen and counted. Later, both of the transactions are concatenated which results in a new dataset. After having a shuffle of this newly created dataset, again the differences are visualized for the reference.

4. LITERATURE REVIEW

Credit Card Fraud can be a life-threating fraud for its users. Most of the web/mobile applications, emerging with new ideas, also incorporate the Online Payment feature for the ease of its users which involves the engagement of credit/debit cards as well. A small mistake of the cardholder may send the invitation to Lawbreakers to access their assets. Although, many researchers are working on Credit Card Fraud Detection System for better results. The literature studied for this paper are summarized as below:

M.R. HaratiNik, M.Akrami, S. Khadivi and M. Shajari [11] had come with a solution of combining Fuzzy expert system and Fogg social investigation hence naming it the FUZZGY half breed model in their research.

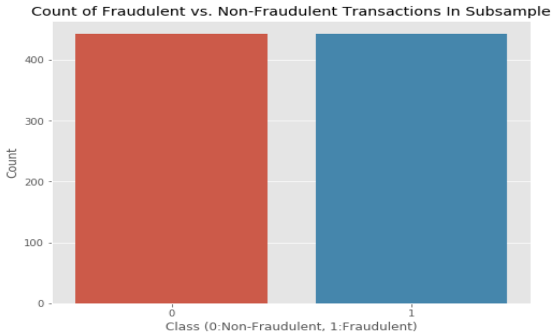


Figure 2: Fraud vs Non-Fraud Transactions in Sample

S. Fashoto, O. Adeleye, and J. Wandera[3] have utilized K-means Clustering algorithm bundling with the Hidden Markov Model (HMM) and Multilayer Perception (MLP) in their research paper. They had made use of K-implies bundling so as to aggregate the speculated deceitful exchanges into a comparable group. The yield of this aspect is utilized to prepare the HMM and the MLP which at that point characterize the approaching exchanges. Their solution recorded the identification exactness of "MLP with K- means Clustering" is higher than the "HMM with K-means Clustering".

Linda Delamaire, Hussein Abdou and John Pointon[1] have come with a solution in which is going to have favorable characteristics with respect to cost investment funds and time productivity. Their research portrays customary terms in card coercion and features key experiences and figures in this field. The important aspects are direct to distinguish the different kind of illicit credit card distortion and inspect elective systems which have been used in deception revelation.

In 2018, Ibtissam Benchaji, Samira Douzi, and Bouabid El Ouahidi[9] has come with an approach to update requested execution of minority of Visa misrepresentation occasions in the unbalanced educational records. To achieve their goal, they proposed a looking system at the subject to the K-suggests gathering and the inherited estimations. They used K-infers figuring to pack and assembling the minority kind of test.

Xuetong Niu, Li Wang, and Xulei Yang have come with an examination where they've worked on Credit Card distortion disclosure by using the distinctive regulated and unsupervised technique in 2019.

In 2009, V. Dheepa, and Dr. R. Dhanapal[2] have presented a research paper which includes three main principles to identify coercion. First, a bundling model was used to arrange the legitimate and phony exchange using data clusterization of regions. Second, the thickness of credit card user's past direct was modeled by the Gaussian Mixture Model with the target that the likelihood of current lead can be set out to perceive any assortments from the standard from the past direct. Eventually, Bayesian frameworks are utilized to delineate the estimations of a specific client and the encounters of various pressure conditions. The standard endeavor is to investigate various perspectives on a comparable issue and see what can be gotten from the utilization of every novel system.

Tanmay Kumar and Suvasini Panigrahi[8] came with a solution in their research paper, in 2015, which projected a mix thanks to agitating credit card extortion recognition utilizing downy bunching and neural system. It utilizes 2 stages. In stage one, they utilized a c-means clustering calculation to provide an incredulous rate of the exchange associate degree in next stage if an exchange is incredulous it's feed into neural system to make your mind up if it had been extraordinarily false or not.

Ayushi Agrawal[12] et al. projected testing associate degree exchange utilizing Hidden mathematician Model, Behavior primarily based strategy and Genetic formula, whereby they utilized the Hidden mathematician Model to stay up the record of past exchanges, Behavior primarily based system for gathering of datasets and ultimately hereditary calculation for advancement for instance computation the limit esteem.

Sam Maes[6] presented police investigation frauds answerable card utilizing two Artificial Intelligence procedures to be specific Bayesian Networks and Artificial Neural Network. The paper examined that however Bayesian systems once brief coaching gave nice outcomes and their speed was improvised by the best use of Artificial Neural Network.

5. PROBLEM FORMULATION

5.1 Dataset (Credit Card Fraud Detection)

The data set used in the proposed work is provided by Kaggle. There are overall 30 features. But 28 out of them are renamed as V1 through V28. All are numeric values. Rest three of the features showcase are the time, calculated amount and whether that transaction was illicit or legitimate. The exact details of the features are hidden for confidentiality. The Class, response variable is 1 for the fraudulent transaction and 0 for safe transactions. The supervised approach is used in this work.

5.2 Algorithm Used

To build Machine Learning models which can distinguish between legitimate and illicit transactions, we've implemented the following models. They are:

- K-Nearest Neighbors (KNN)
- Logistic Regression
- Random Forest
- XGBoost (Extreme Gradient Boost)

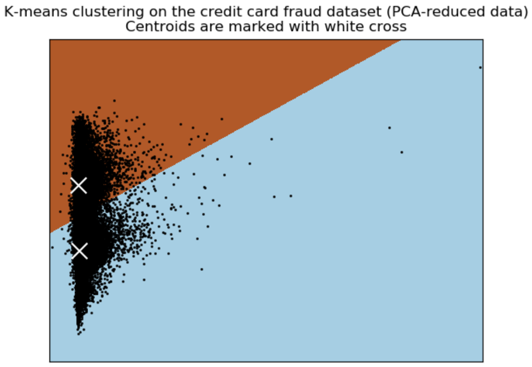


Figure 3:K-Means Clustering

5.2.1 K Nearest Neighbors (KNN)

A straight forward algorithm capable of collecting all the possible cases, measuring the similarities and classifying new cases based on homogenous measurement which helps to in identifying patterns and statistical estimations. Similarities, in this algorithm, is defined by the distance matric in between of two data points x and y. The distance matric can be formulated as below:

$$\mathbf{d(x, y)} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots\ldots\ldots\ldots + (x_n - y_n)^2} \qquad (1)$$

Suppose, A is the set of Kpoints in the trainingdataset, which is closest to x where, for each class, conditionalprobability can be formulatedas the portion of marks in A with that given class label.

$$\text{P(y=j | X=x)} = \frac{1}{K}\sum_{i \in A} I(y^i = j) \qquad (2)$$

Where the value of x is treated as true when 1 is returned by the I(x), aka indicator method, otherwise false. Eventually, the input x is committed to the class with the highest expectation.

5.2.2 *Logistic Regression*

Logistics Regression is another statistic technique acquired by the Machine Learning to scrutinize the dataset when the dependent parameter is dichotomous i.e. the data can only in form of 1(True, Yes, pregnant, etc) or 0(False, No, non-pregnant, etc). It was named for the sigmoid/logistic function used at the core of the method. The intention of *Logistic regression* is to return the best modelwhich can be used to determinethe dataset and to depict the relation between one dependent dichotomous variable and a portion of ordinal, nominal, interval or ratio-level independent variables.

Logistic Regression formulates the coefficient to predict logit function which can be given by:

$$Logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots\dots\dots + \beta_k X_k \quad (3)$$

Where p is the probability of characteristics of interests.

Logit function as logged odds where odds are the fraction of probabilities of presence and absence of characteristics:

$$\text{odds}=\frac{p}{1-p}=\frac{probability\ of\ presence\ of\ characterstics}{probability\ of\ absence\ of\ characterstics} \quad (4)$$

$$\text{and logit(p) = Ln}\left(\frac{p}{1-p}\right) \quad (5)$$

5.2.3 *Random Forest*

The random forest algorithm by L. Breiman 2001, has been successful as a general-purpose classification and regression method. This approach uses several randomized decision trees and aggregates their predictions by averaging, has shown excellent performance in the setting where the number of observations is very less in comparison to the number of variables. This solution is applied to dominant problems whichare easily adopted to various ad-hoc learning tasks,and returns measures of variable importance. Random Forest can be applied for both classification and regression situations where the dependent variable is categorical in regression whereas continuous in classification.

Approach of Random Forest : In Random Forest, randomized decision trees are created by the help of Bagging Algorithm. An n by m dimensional dataset is taken and from those one third was left out which is known as Out of the Bag Sample. This data then used to analyze the unbiased calculate of the error. A new randomized dataset is subsetted and trained from the original dataset (two-third part) for sampling x number of cases by replacing the original dataset. From the m columns, M are selected randomly for each node where M is m/3 for regression and the square root of M for classification. There is no pruning in Random Forest tree i.e. it growns fully. Pruning means choosing a subtree which must lead to the least error rate. Several trees are grown and the final prediction is obtained by calculating mean of all.

5.2.4 *XGBoost*

XGBoost model is also compared, which is based on Gradient Boosted Trees and is a more powerful model compared to both Logistic Regression and Random Forest.

Below illustrated characteristics make XGBoost most popular which overcome the limitation of other modeling techniques as well.

- Speed and Performance
- Parallelizability
- Consistently outperforms other modeling techniques
- Numerous configurable tuning options

XGBoost comes under the family of Boosting that operates on the principle of the ensemble. It works on the methodology of combining a set of weak learners and outcomes with improved prediction. Boosting can be best defined with the below scenario.

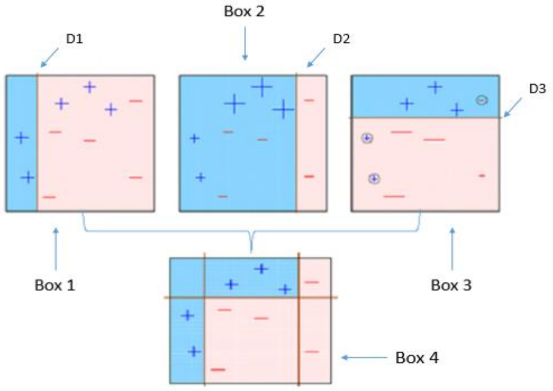


Figure 4: Boosting Explained

As in the above figure, it can be seen that the first decision stump (D1) is divided into two regions – the blue region (+) and red region (-). Also, the red region has three incorrectly classified (+) which weights more than other observations and inputs to the second learner. This data modeling continues and regulates the error faced by last observation until most accurate prediction model outcomes.

5.3 *Performance Analysis*

We've studied the different models for the detection of Credit Card Fraud. Performance of each algorithm can be seen by the performance table below:

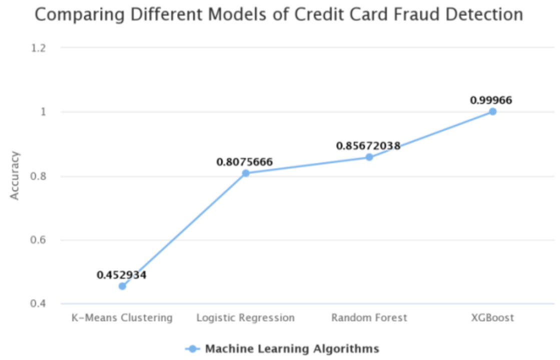


Figure 5: Comparison chart of Prediction Accuracy of Different Models

6. CONCLUSION

The proposed methodology adopted is proficient and viable. We had the option to precisely recognize fraudulent credit card transactions utilizing the random forest model and XGBoost. The fraud transactions can look fundamentally the same as standard exchanges, it is hard to place them into a different gathering dependent on highlights alone. The K-implies grouping model delivered a low precision of 54.27%. Subsequently, K-means would not be the favored model for this data set, as it didn't effectively anticipate cheats and it likewise created plenty of false positives. The strategic relapse gave us the best outcomes. The logistic regression gave us an extraordinary precision rate of 99.88%, with 0.079% of the approval set being false negatives (or 0.49% of the number of frauds). It has appeared even a logistic regression model can accomplish great review, while a significantly more complex Random Forest model enhances strategic relapse as far as AUC. Be that as it may, the XGBoost model enhances the two models. Improvement of the random forest model is possible by further manipulating the hyperparameters, given additional time and/or computational power.

7. FUTURE SCOPE

Credit Card Fraud Detection is a mind-boggling matter where a considerable measure of arranging is essential prior to toss the AI algorithms. In any case, it is likewise a utilization of information science and AI for the great, which ensures that the cash of the customers safe and not effectively messed with. In future work, Random forest mode would be improved for detecting fraudulent transactions.

REFERENCES

[39] Linda Delamaire, Hussein Abdou and John Pointon(2009) ”Credit card fraud and detection techniques: a review” Banks and Bank Systems, Volume 4, Issue 2, 2009

[40] V. Dheepa, and Dr. R.Dhanapal (2009) “Analysis of Credit Card Fraud Detection Methods” International Journal of Recent Trends in Engineering, Vol 2, No. 3, November 2009

[41] O. O. O. A., W. Stephen Fashoto, “Hybrid Methods for credit card fraud detection,” Kampala International University, Kampala, Uganda, University of Abuja, Nigeria, Redeemer's University, Ede, Osun State, Nigeria, [Online]. Available: http://www.journalrepository.org/media/journals/BJAST_5/2015/Dec/Fashoto1352015BJAST21603.pdf.

[42] Zareapoor M, Seeja KR, Alam AM. Analyzing credit card: fraud detection techniques based on certain design criteria. International Journal of Computer Application. 2012;52(3):35–42.

[43] Duman E, Ozcelik MH. Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications. 2011;38(10):13057–13063.

[44] Maes S, Tuyls K, Vanschoenwinkel B, Manderick B. Credit card fraud detection using Bayesian and neural networks. Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies; 1993; pp. 261–270.

[45] Srivastava A, Kundu A, Sural S, Majumdar AK. Credit card fraud detection using hidden Markov model. IEEE Transactions on Dependable and Secure Computing. 2008;5(1):37–48.

[46] Panigrahi S, Kundu A, Sural S, Majumdar AK. Credit card fraud detection: a fusion approach using Dempster-Shafer theory and Bayesian learning. Information Fusion. 2009;10(4):354–363

[47] Ibtissam Benchaji, Samira Douzi, and Bouabid El Ouahidi(2018) “Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for credit card fraud detection” 2nd Cyber Security in Networking Conference (CSNet)

[48] [Zareapoor and Shamsolmoali 2015] Zareapoor, M., and Shamsolmoali, P. 2015. Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia Computer Science48:679–685.

[49] M. A. S. K. M. S. MR HaratiNik, “FUZZGY model,” [Online]. Available: https://ieeexplore.ieee.org/document/6483148.

[50] S. k. A. K. M. Ayushi agarwal, “Credit card fraud detection: A case study,” published in IEEE, New Delhi, India, 2015.