JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 2, 2020

COMPARING DIFFERENT MODELS FOR CREDIT CARD FRAUD DETECTION

Vinaya Keskar

Research Student at Savitribai Phule Pune University, Pune
Asst. Professor, ATSS's College of Business Studies and Computer Applications, Pune.

vasanti.keskar@gmail.com

ABSTRACT: Credit Card Fraud detection is a challenging task for researchers as fraudsters are innovative, quick-moving individuals. Credit Card fraud detection system is challenging as the dataset provided for credit card fraud detection is very imbalanced. 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 research aims to enhance the performance of the minority of credit card fraud on the dataset available. So, K-means clustering, logistic regression, random forest and XG Boost models are performed. This research work incorporates Credit Card Fraud Detection models to study the transactions that end with some frauds. This paper is then used to distinguish whether payment transactions are fraud or not. This research work is to identify false transactions totally while avoiding incorrect fraud classifications. Different algorithms are implemented in this paper. Python Machine Learning libraries are used to perform those algorithms. The models studied in this research work are K-Nearest Neighbor, logistic regression, random forest model, XG Boost model. XG Boost is showing more accuracy than other models. Out of these algorithms, the XG Boost model is preferable over the Random Forest model and Logistic regression model.

KEYWORDS: Credit Card Fraud Detection, K-Means Clustering, Logistic Regression, Random Forest, XG Boost Model

I. INTRODUCTION

Credit Card Fraud detection is a challenging task for researchers as fraudsters are innovative, quick-moving individuals. This Research is done to prevent the problem of society. With the normal procedure, it is not possible to perform fraud. The scamster must do their work with great confidence and sharply. So, businesses, as well as academic communities, are developing credit card fraud detection models. In this study, different approaches to fraud detection are presented. This study investigates the usefulness of applying different approaches to a problem of Credit card fraud detection.

The principle kind of credit card extortion is the illegal use of Lost and Stolen Cards. Creating Fake and Doctored Cards is in cutting-edge procedures. The data is hung on either the attractive strip on the rear of the credit card or the data set aside on the savvy chip is copied beginning with one card then onto the following. Phishing websites are changing into a standard method of blackmail with a talented hacking limit. These pages are proposed to get individuals to give their credit-cards details effectively without recognizing they have been frauded. Triangulation is comparably another method. Triangulation is the point at which a seller offers a thing at an incredibly unassuming cost through a site. When a client attempts to submit the item's request, the transporter urges to the customers or clients who are insnared, to pay by methods for the email if he gets the item. The seller uses misleading card details for buying the item from a site and sends the thing to the buyer, who sends the trader their credit card details using email. The trader continues working along these lines utilizing the credit-card numbers given by buyers for buying; showing up for a brief time frame to be the individual closes the web page and begins another one.

Credit-card misrepresentation can be done in two ways, either offline or online misrepresentation (fraud). The possibilities of offline fraud have a stolen physical card at client confronting venue or by the call center. The card-issuing banks or organizations can jolt it before it is used in a bogus manner. Online misrepresentation is submitted through phone shopping, web or when cardholders not present. The loss of individual data straightforwardly adds to developing misrepresentation misfortunes for banks and merchants. It is essential to collect key details for breaches to avoid losses from a financial service organization. This information is planned to help extortion supervisory groups figure out where holes exist in this industry's security issues.

II. RELATED WORK

[1] Portrayed highlights key experiences and customary terms in Credit card coercion and figures in this type of fraud. Dependent upon the distortion looked by credit card companies or banks; various counters measures can

ISSN- 2394-5125 VOL 7, ISSUE 2, 2020

be gotten and completed. The proposals made were presumably going to have favorable characteristics regarding cost investment funds and time productivity. [2] Three procedures to recognize coercion were presented. At first, the gathering model was used to arrange the legal and phony exchange using data clustering of parameter regard areas. Also, Gaussian blend model was utilized to show the likelihood thickness of credit card. Client's past direct with the target that the likelihood of current lead can be set out to perceive any assortments from the standard from the past direct. Finally, Bayesian frameworks were utilized to delineate a specific client's estimations and the encounters of various pressure conditions. The standard endeavour was to investigate various perspectives on a comparable issue and see what can be gotten from every novel system's utilization. [3] Used machine learning AI methods for data mining to recognize extortion in a progressing exchange on the web and organize the exchange of information as genuine or suspicious. [4] They have implemented different extortion recognition frameworks such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Hidden Markov Model, Artificial Neural Networks (ANN), Bayesian Network, Fuzzy Logic Based System and Decision Trees. According to the requirement, a fair audit was done on the current and proposed models for credit card extortion discovery and completed a near report on these strategies based on quantitative estimations, such as exactness, recognition rate and false alert rate. The end of the examination clarifies the downsides of existing models and gives superior security arrangements.

III. DATASET

Kaggle provides the data set used in the proposed work. There are overall 30 features. But 28 out of them are renamed as V1 through V28. All are numeric values. The remaining three features are the time, transaction amount status of the transaction, it was fraudulent or not. 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. There are no missing values in the data set.

IV. IMPLEMENTATION

System Configuration

In this research, all the tests are performed under following specifications:

Host System: Intel i3 processor with 4 GB RAM and 500GB Hard disk.

Operating Environment: Windows10

IDE: Visual StudioCode

Programming Language: Python (v3.6)

Algorithms Implemented

To predict whether a transaction is fraudulent, machine learning models are implemented. The following models are implemented.

- K Nearest Neighbors (KNN)
- Logistic regression
- Support vector classifier
- Random forest
- XG Boost

K Nearest Neighbors

In the classification setting, the KNN algorithm is forming a majority vote between the K most similar instances to a given unseen observation. Similarity is defined according to a distance metric between two data points x and x_0. A popular choice is the Euclidean8distance given by

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + ... + (x_n - x_n')^2}$$

Suppose the K points in the training data that is closest to x are denoted as set A. It then estimates the conditional probability for each class, i.e., the fraction of points in A with that given class label.

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

Where I(x) is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise finally, the input x is assigned to the class with the largest probability.
Output generated by proposed work:

ISSN- 2394-5125 VOL 7, ISSUE 2, 2020

```
Accuracy: 0.4529349803696256
False negative rate (with respect to misclassifications):  0.0010696851235972539
False negative rate (with respect to all the data):  0.0005851873131390512
False negatives, false positives, mispredictions:  55 51362 51417
Total test data points: 93987
```

Logistic Regression and Support Vector Classifier

Logistic regression introduced by David Cox, 1958 is a model in which the response variable Y is categorical. SVM, introduced by Vapnik, 1995, was to solve the classification and regression problems. It is used to implement an optimal hyper plane for maximizing the margin between two classes.

To learn much faster on large datasets, class SGD Classifier implements SGD training with multiple linear classifiers.

Output: 0.8075666039570759

We have 492 fraud data points and 284315 non-fraudulent data points.

And the generated output is:

```
X and y sizes, respectively: 284807 284807
Train and test sizes, respectively: 185124 185124 | 99683 99683
Total number of frauds: 492 0.001727485630620034
Number of frauds on y_test: 188 0.0018859785520098714
Number of frauds on y_train: 304 0.0016421425639031135
Accuracy:  0.9989165655126752

Confusion matrix:

Predicted  False     True     __all__
Actual
False      99463     32       99495
True       76        112      188
__all__    99539     144      99683
```

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 having greater the number of observations then the number of variables.

Output: 0.8567203839327452

As compared to SVC, random forest performed very well.

XG Boost

XG Boost model is also compared based on Gradient Boosted Trees and is a more powerful model compared to Logistic Regression and Random Forest. XG Boost uses the gradient boosting (GBM) framework at its core. It uses the features of scikit-learn library. It is an optimized, distributed gradient boosting library. But wait, what is boosting. XG boost is effective for data in tabular form having limited variables set.

TEST SET EVALUATION OF THE BEST MODEL

According to the cross-validated MCC scores, the best-performing model is random forest, evaluating its performance:

ISSN- 2394-5125 VOL 7, ISSUE 2, 2020

For Random Forest Model

```
CONFUSION MATRIX
[[56854    10]
 [   15    83]]

CLASSIFICATION REPORT
             precision    recall  f1-score   support

          0    0.99974   0.99982   0.99978     56864
          1    0.89247   0.84694   0.86911        98

avg / total    0.99955   0.99956   0.99956     56962

SCALAR METRICS
            MCC = 0.86919
          AUPRC = 0.85098
          AUROC = 0.95924
Cohen's kappa = 0.86889
       Accuracy = 0.99956
```

For XG Boost Model

```
Confusion Matrix and Statistic

          Reference
Prediction  0    1
     0.56863    17
     1     6    75

               Accuracy: 0.9996
               95% CI: (0.9994, 0.9997)
No information rate: 0.9984
P-Value [Acc > NIR] : < 2e-16
```

ISSN- 2394-5125 VOL 7, ISSUE 2, 2020

Kappa: 0.8669

Mcnemar's Test P-value: 0.03706

Sensitivity: 0.9999

Specificity: 0.8152

Pos Pred Value: 0.997

Neg Pred Value: 0.9259

Prevalence: 0.9984

Detection rate: 0.9983

Detection Prevalence: 0.9986

Balanced Accuracy: 0.9076

'Positive' Class: 0

According to the MCC, the performance of random forest is not so effective on the training set.

V. CONCLUSION

The proposed methodology adopted is proficient and viable. The random forest model and XG Boost are options to recognize fraudulent credit card transactions precisely. We found that the five factors V17, V14, V10, V12, and V11, are most associated with the fraud. 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 a great deal 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 XG Boost model enhances the two models. The Random forest model can be improved further by manipulating the hyper parameters.

VI. FUTURE SCOPE

Issues of Fraud detection are mind-boggling require a considerable measure before the implementation of AI algorithms. The utilization of AI algorithms and business analytics ensures that the customers' cash will be safe and not effectively messed with. In future work, Random forest mode would be improved for detecting fraudulent transactions.

VII. REFERENCES

[1] 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

[2] 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

[3] John Akhilomen(2013) "Data Mining Application for Cyber Credit-Card Fraud Detection System" Advances in Data Mining. Applications and Theoretical Aspects pp 218-228

[4] Excell D. Bayesian inference—the future of online fraud protection. Computer Fraud and Security. 2012; 2012(2):8–11.

[5] Lu Q, Ju C. Research on credit card fraud detection model based on class weighted support vector machine. Journal of Convergence Information Technology. 2011; 6(1):62–68.

[6] Juszczak P, Adams NM, Hand DJ, Whitrow C, Weston DJ. Off-the-peg and bespoke classifiers for fraud detection. Computational Statistics and Data Analysis. 2008; 52(9):4521–4532.

[7] Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: a comparative

study. Decision Support Systems. 2011; 50(3):602–613.

[8] Francisca NO. Data mining application in credit card fraud detection system.Journal of Engineering Science and Technology. 2011; 6(3):311–322.

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

[10] 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.

[11] Chan PK, Fan W, Prodromidis AL, Stolfo SJ.Distributed data mining in credit card fraud detection.IEEE Intelligent Systems and Their Applications. 1999;14(6):67–74.]

[12] N. Sivakumar and Dr.R.Balasubramanian(2015) "Fraud Detection in Credit Card Transactions:Classification, Risks and Prevention Techniques" International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1379-1386

[13] Ibtissam Benchaji, SamiraDouzi, 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)

[14] Yashvi Jain, NamrataTiwari, Shripriya Dubey and Sarika Jain (2019) "A Comparative Analysis of Various Credit Card Fraud Detection Techniques" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-5S2, January 2019

[15] R. M. jamailesmaily, "Intrusion detection system based on multilayer perceptron neural networks and decision tree," in International conference on Information and Knowledge Technology, 2015.

[16] S.P. Tanmay Kumar Behera, "Credit card fraud detection: a hybrid approach using fuzzy clustering and neural network," in international conference on advances in computing and communication Engineering, 2015.

[17] P. K. D. K. R. D. A. A. Thuraya Razoogi, Credit card fraud detection using fuzzy logic and neural networks, Society for modelling and simulation International(SCS), 2016.

[18] Zojaji et al. 2016] Zojaji, Z.; Atani, R. E.; Monadjemi,A. H.; et al. 2016. A survey of credit card fraud detection techniques: data and technique-oriented perspective.