IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906251 |

Credit Card Fraud Detection Using Machine Learning Algorithms

Sonali Kadam, Shristy Kumari, Subhu Trivedi, Vanshika Shah

Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women,

Pune, Maharashtra, India

ABSTRACT: Due to a rapid advancement in the electronic commerce technology, the use of credit cards has dramatically increased. E-commerce and many other online sites have increased the online payment modes. Since, credit card is one of the most popular modes of payment, the number of fraud cases associated with it is also rising. Thus, in order to stop these frauds, we need a powerful fraud detection system that detects it in an accurate manner.
In this paper we have explained the concept of frauds related to credit cards. Here we implement different machine learning algorithms on an imbalanced dataset such as logistic regression, support vector machine, random forest.

KEYWORDS: Credit card, PCA (Principal Component Analysis), Machine Learning.

I. INTRODUCTION

Computer and Internet are now playing a major role in every aspect of human lives which mainly includes data, transactions, information storage and its retrieval. As we all are living in the 20th century, everything is being performed online like online shopping, bill payments etc. through various online platforms like websites, android applications etc. which has led us to move towards the digital era. Digitalization is basically the use of digital technologies to change a business model and provide new revenue and value-producing opportunities; it is the process of moving to a digital business. Thus, increase in the digitalization has led to the increase of fraudulent activities in various sectors. There are various types of frauds which occurs on online platform and one such attack is credit card frauds.

II. RELATED WORK

Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule or policy with an aim to attain unauthorized financial benefit. Numerous literatures pertaining to anomaly or fraud detection in this domain have been published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates have revealed that techniques employed in this domain include data mining applications, automated fraud detection, adversarial detection.

Multiple Supervised and Semi-Supervised machine learning techniques are used for fraud detection [8], but we aim is to overcome three main challenges with card frauds related dataset i.e., strong class imbalance, the inclusion of labelled and unlabelled samples, and to increase the ability to process a large number of transactions. Different Supervised machine learning algorithms [3] like Decision Trees, Naive Bayes Classification, Least Squares Regression, Logistic Regression and SVM are used to detect fraudulent transactions in real-time datasets.

An Artificial Immune Recognition System (AIRS) for credit card fraud detection was proposed in [7]. AIRS is an improvement over the standard AIS model, where negative selection was used to achieve higher precision. A credit card fraud detection system was proposed in [5], which consisted of a rule-based filter, Dumpster–Shafer adder, transaction history database, and Bayesian learner. The Dempster–Shafer theory combined evidential information and created an initial belief, which was used to classify a transaction as normal, suspicious, or abnormal. If a transaction was suspicious, the belief was further evaluated using transaction history from Bayesian learning [5].

A hybrid clustering system with outlier detection capability was used in [18] to detect fraud in lottery and online games. The system aggregated online algorithms with statistical information from the input data to identify a number of fraud types. The training data set was compressed into the main memory while new data samples could be incrementally added into the stored data-cubes. The system achieved a high detection rate at 98%, with a 0.1% false alarm rate [18].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906251 |

III. CHALLENGES IN CREDIT CARD FRAUD DETECTION

A. *BehavioralVariation:*

Fraudulent behaviour tends to change over time in order to avoid detection. Credit card fraud detection should not be static i.e., constructed once and never updated. Known methods used for overcoming concept drifts problem are adaptive based learners and ensemble. Adaptive based learners have a drift detection mechanism that updates the current model when the drift is detected while ensembles have natural ability to retain relevant information and acquire knowledge.

B. *Cost Sensitive Problem:*

Credit card fraud detection is a cost sensitive problem which means that the cost produced by misclassifying genuine transaction is different than the cost of misclassifying fraudulent one. Failure to detect a fraudulent transaction causes the financial loss of the amount of that transaction. The problem occurs because of the overlapping data – when many genuine transactions resemble fraudulent one and vice versa.

C. *Imbalanced Data:*

On the global level, fraudulent transactions are amounted to less than 0.05% of the total transactions. If this problem had not been taken into consideration any machine algorithm that classifies correctly only genuine transactions would perform outstanding, with the accuracy level above 99%, disregarding the fact that all the minority class transactions are classified falsely. Data level methods, such as over sampling and under sampling, alter the size of dataset used for training. While the level of imbalance is reduced problems of over-fitting ignoring useful data are prevalent.

Complex sampling method SMOTE, oversamples the minority class generating synthetic examples by interpolating k minority class nearest neighbours. Through this process the classifier builds larger decision region that contain nearby examples from minority class, which has shown improvements in application.

D. *Data Deficiency:*

Essentially, the biggest problem in dealing with credit card fraud detection scientifically is that real data hardly ever available for exploration, due to the issue of confidentiality. But the researchers can still carry out the scientific work by associating with the respective industrial partner, who provides the data. Also, synthetic data which simulates dataset of transactions can be another option.

IV. ALGORITHM

The dataset [11] contains transactions made by a cardholder in the month of September 2013. Where there is total 284,807 transactions among which there are 492 i.e., 0.172% transactions are fraudulent transactions. This dataset is highly unbalanced. Since providing transaction details of a customer is considered to issue related to confidentiality, therefore most of the features in the dataset are transformed using principal component analysis (PCA). V1, V2, V3..., V28 are PCA applied features and rest i.e., 'time', 'amount' and 'class' are non-PCA applied features.

Based on the previous research, we used three algorithms that are among five most used in credit card fraud detection: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR).

A. *Support Vector Machine (SVM):*

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. After that, we perform classification by finding the hyper-plane that differentiates the two classes very well. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

B. *Logistic Regression:*

Logistic Regression has its history in fraud classification. Logistic Regression is probabilistic model – it assigns probability to each classified sample. This gives us more possibilities than pure binary classification. Interims of credit card fraud detection, this allows us to rank the transactions by their probability to be fraudulent and choose the threshold for most probable frauds that will raise the alert by predictive model.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906251 |

C. *Random Forest:*

Random forests are the classifiers that combine many tree possibilities, where each tree depends on the values of a random vector sampled independently then, all trees in the forest will have same allotment. To construct a tree, we assume that n is the number of training observations and p is the number of variables (features) in a training set. To determine the decision node at a tree we choose k « p as the number of variables to be selected. We select a bootstrap sample from the n observations in the training set and use the rest of the observations to estimate the error of the tree in testing phase. Hence, we randomly choose 'k' variables as a decision at certain node in the tree and calculate the best split based on the k variables in the training set. Trees are always grown and never pruned compared to other tree algorithms. Random forests can handle large number of variables in a data set. Also, during the forest building process they generate an internal unbiased estimate of the generalization error. Additionally, they can estimate missing data closely. A major disadvantage of random forests algorithm is it does not give precise continuous forecast.

V. EVALUATION AND RESULTS

The dataset used is the only publicly dataset that is available for credit card fraud detection. There is total 284807 transactions out of which 492 are fraudulent ones. The dataset is highly imbalanced, as it has only 0.172% fraudulent transactions. To handle class imbalance, we can choose SMOTE method. PCA can be used for dimensionality reduction.

The evaluation of the results can be done by two measures, average precision and area under the ROC curve (AUC). Average Precision approximates area under the precision-recall curve. A precision-recall curve is a plot of precision versus recall at different probability thresholds. ROC curve is plotted as recall (TPR) against Fall-out (FPR) at various classification thresholds. Transaction with fraud probability that equals or is above threshold value is considered fraudulent. The best classifier corresponds to the point (0,1) where there are no false positives or false negatives. AUC metric measures how much is the ROC curve of single classifier close to the optimal point.

VI. CONCLUSION

A study on credit card fraud detection using machine learning algorithms has been presented in this paper. Also the challenges in credit card fraud detection are mentioned. We have studied and measured performance of three selected ML algorithms: Random Forest, Support Vector Machine and Logistic Regression and shows that it proves accurate in deducting fraudulent transaction and minimizing the number of false alerts. Comparison of all the three methods can be done to find out which algorithms gives more accurate results. If these algorithms are applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions.

REFERENCES

1. N. Mahmoudi and E. Duman, ''Detecting credit card fraud by modified fisher discriminant analysis,'' Expert Syst. Appl., vol. 42, no. 5, pp. 2510–2516, 2015.
2. Y. Sahin, S. Bulkan, and E. Duman, ''A cost-sensitive decision tree approach for fraud detection,'' Expert Syst. Appl., vol. 40, no. 15, pp. 5916–5923, 2013.
3. Mohammed, Emad, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1 July 2018, doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.
4. E. Duman and M. H. Ozcelik, ''Detecting credit card fraud by genetic algorithm and scatter search,'' Expert Syst. Appl., vol. 38, no. 10, pp. 13057–13063, 2011.
5. S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, ''Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning,'' Inf. Fusion, vol. 10, no. 4, pp. 354–363, 2009.
6. Randhawa, Kuldeep, et al. "Credit Card Fraud Detection Using AdaBoost and Majority Voting." IEEE Access, vol. 6, 2018, pp. 14277–14284., doi:10.1109/access.2018.2806420.
7. N. S. Halvaiee and M. K. Akbari, ''A novel model for credit card fraud detection using artificial immune systems,'' Appl. Soft Comput., vol. 24, pp. 40–49, Nov. 2014.
8. Melo-Acosta, German E., et al. "Fraud Detection in Big Data Using Supervised and Semi-Supervised Learning Techniques." 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 2017, doi:10.1109/colcomcon.2017.8088206.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906251 |

9. "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018.
10. Lakshmi S V S S, Selvani Deepthi Kavila, Machine learning for credit card fraud detection system, International Journal Of Applied Engineering Research ISSN 2018.
11. J. T. Quah and M. Sriganesh, ''Real-time credit card fraud detection using computational intelligence,'' Expert Syst. Appl., vol. 35, no. 4, pp. 1721–1732, 2008.
12. C.-F. Tsai, ''Combining cluster analysis with classifier ensembles to predict financial distress,'' Inf. Fusion, vol. 16, pp. 46–58, Mar. 2014.

BIOGRAPHY

Subhu Trivedi is a final year computer year engineering student (batch 2020-21) from Bharati Vidyapeeth's College of Engineering for Women, Pune which is affiliated to Savitribai Phule Pune University (formerly known as University of Pune).

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details