IJIRMET ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering and Technology
Vol. 6, Issue 4, April 2021

CREDIT CRD FRAUD DETECTION Using SUPERVISED MCHINE LEARNING ALGORITHM:RANDOM FOREST

[1] D. Rajapriya,[2] Fathima Nezlin. N

[1] Head Of The Department

[1][2] Department of Computer Science And Engineering, RVS Technical Campus, Coimbatore-641402, Anna University, Chennai, India.

Abstract: ***Credit card fraud events take place frequently and then result in huge financial losses. Criminals can use some technologies such as Trojan or Phishing to steal the information of other people's credit cards. Therefore, an effective fraud detection method is important since it can identify a fraud in time when a criminal uses a stolen card to consume. One method is to make full use of the historical transaction data including normal transactions and fraud ones to obtain normal/fraud behavior features based on machine learning techniques, and then utilize these features to check if a transaction is fraud or not. In this paper, Machine Learning algorithm is used to train the behavior features of normal and abnormal transactions. We implement this using Random forest machine learning algorithm in Open CV and analyze the performance on credit fraud detection.***

1. INTRODUCTION

Nowadays Credit card usage has been drastically increased across the world, now people believe in going cashless and are completely dependent on online transactions. The credit card has made the digital transaction easier and more accessible. A huge number of dollars of loss are caused every year by the criminal credit card transactions. Fraud is as old as mankind itself and can take an unlimited variety of different forms. The PwC global economic crime survey of 2017 suggests that approximately 48% of organizations experienced economic crime. Therefore there's positively a necessity to unravel the matter of credit card fraud detection. Moreover, the growth of new technologies provides supplementary ways in which criminals may commit a scam. The use of credit cards is predominant in modern day society and credit card fraud has been kept on increasing in recent years. Huge Financial losses have been fraudulent effects on not only merchants and banks but also the individual person who are using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that. For example, if a cardholder is a victim of fraud with a certain company, he may no longer trust their business and choose a competitor.

Fraud Detection is the process of monitoring the transaction behavior of a cardholder to detect whether an incoming transaction is authentic and authorized or not otherwise it will be detected as illicit. In a planned system, we are applying the random forest algorithm for classifying the credit card dataset. Random Forest is an associate in the nursing algorithmic program for classification and regression. Hence, it is a collection of decision tree classifiers. The random forest has an advantage over the decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. The paper focuses on the use of Regression and LSTM based Machine learning to predict stock values. Factors considered are open, close, low, built, each node then splits on a feature designated from a random subset of the complete feature set. Even for large data sets with many features and data instances, training is extremely fast in the random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to overfitting. There are various fraudulent activities detection techniques has implemented in credit card transactions have been kept in researcher minds to methods to develop models based on artificial intelligence, data mining, fuzzy logic and machine learning. Credit card fraud detection is a very troublesome, but also a popular problem to solve. In our proposed system we built the credit card fraud detection using Machine learning.

With the advancement of machine learning techniques. Machine learning has been recognized as a no-hit live for fraud detection. A great deal of data is transferred throughout on-line transaction processes, resulting in a binary result: genuine or fraudulent. Online businesses are able to identify fraudulent transactions accurately because they receive chargebacks on them. Within the sample fraudulent datasets, features are constructed.
These area unit information points like the age and price of the client account, as well as the origin of the credit card. There are many options and everyone contributes, to varying extents, towards the fraud probability.
Note, the degree within which every feature contributes to the fraud score isn't determined by a fraud analyst, but is generated by the artificial intelligence of the machine which is driven by the training set. So, in regard to the card fraud, if the use of cards to commit fraud is proven to be high, the fraud weighting of a transaction that uses a credit card will be equally so. However, if this were to diminish, the contribution level would parallel. Simply put, these models self- learn while not express programming like with manual review.
Credit card fraud detection using Machine learning is done by deploying the classification and regression algorithms. We use a supervised learning algorithm such as Random forest algorithm to classify the fraud card transaction online or by offline. Random forest is an advanced version of the Decision tree. The random forest has better efficiency and accuracy than the other machine learning algorithms. Random forest aims to reduce the previously mentioned correlation issue by choosing only a subsample of the feature space at each split. Essentially, it aims to make the trees de-correlated and prune the trees by setting a stopping criterion for node splits.

I. LITERATURE SURVEY

[1] Number of literature pertaining techniques to detect frauds or anomalies regarding the credit card transactions have been already published and are available for public usage. We are using Random Forest as our base classifier for classification of fraudulent and legitimate credit card transactions. Similarly The Random forest as a base classifier has been used by Shiyang Xuan et.al.

[2] which introduces two methods, first one is Random-Tree-based Random Forest which is a simple implementation of decision trees in which at each internal node a subset of attributes are selected randomly and classify data sets into two classes namely fraudulent and legitimate transaction, another one is CART based Random Forest in which at each node data sets are split by choosing the best attribute from the subset of attributes. The random forest generated by this method must have some technique to limit the growth of decision trees. The voting machine which has been used in the proposed model assumes that each base classifier has equal weight but some may be more important than others.Such voting machine method is implemented our proposed model. But it has been found by Kuldeep Randhawa that instead of majority voting techniques, AdaBoost gives better performance

[3].A hybrid model is used while using this technique which increases the training time of the model which is not suitable for online learning. While using Random Forest as a base classifier there is need of limiting the growth of decision trees in random forest. This work has been efficiently done by Angshuman Paul et.al.

[4].An improved classifier has been built which performs better with the minimum number of trees. Here the thing which is to be focused is that in the proposed method number of trees is not predefined i.e. low data dependence. Also by limiting the growth of the tree makes the method fast and useful for industrial applications. Though the method performs well than classic Random Forest Classifier, it has been also found that there are other classification algorithms like neural networks which still outperforms the proposed method. The Neural network and Whale algorithm has been used by Chuzi wang et.al. adopts the advantages of BP neural network algorithm and whale algorithm and gives high detection accuracy and fast convergence speed

[5].Thus, this model have higher accuracy than the improved version of random classifier which is discussed earlier. But as model have higher accuracy, it is more prone to over-fitting problem which is need to be handled properly. Krishna Modi et.al. done analysis and compare the neural network technique with HMM and Decision

EXISTING SYSTEM AND ITS DRAWBACKS

In existing System, a research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods

for fraud detection and to increase the accuracy of results. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential. Accuracy of an algorithm is around 50%. Significance of this paper was to find an algorithm and to reduce the cost measure. The result obtained was by 23% and the algorithm they find was Bayes minimum risk.

Demerits of Existing System:

1. In this paper a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed.
2. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure.

PROPOSED SYSTEM AND ITS ADVANTAGES

We are applying random forest algorithm for classify the credit card dataset. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction[8]. Random Forest is an algorithm for classification and regression. It is a collection of decision tree classifiers. Random Forest adds additional randomness to the model searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that results in a better model. Random forest has precedence over decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each single tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. By using Random Forest algorithm the generalization error and over fitting were achieved.
In classification this is the mode (or most common) class value. Its purpose is to used database in which the data points are separated into several classes to predict the classification of a new sample point . Classification steps

Training phase: a model is constructed from the training instances. classification algorithm finds relationship between predictors and targets.

Testing phase: test the model on a test sample whose class labels are known but not used for training the model.

Usage phase: use the model for classification on new data whose class labels are unknown

Merits of Proposed System:

- Random forest ranks the importance of variables in a regression or classification problem in a natural way can be done by Random Forest.
- The 'amount' feature is the transaction amount. Feature 'class' is the target class for the binary classification and it takes value 1 for positive case (fraud) and 0 for negative case (not fraud).

CONCLUSION

Hence, we can acquired the result of an accurate value of credit card fraud detection i.e. 0.9994802867383512 (99.93%) using a random forest algorithm with new enhancements. In comparison to existing modules, this proposed module is applicable for the larger dataset and provides more accurate results.
The Random forest algorithm will provide better performance with many training data, but speed during testing and application will still suffer. Usage of more pre-processing techniques would also assist. Our future work will try to represent this into a software application and provide a solution for credit card fraud using the new technologies like Machine Learning, Artificial Intelligence and Deep Learning.

FUTURE ENHANCEMENT

While we couldn't reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This

provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

Our future work will try to represent this into a software application and provide a solution for credit card fraud using the new technologies like Machine Learning, Artificial Intelligence and Deep Learning.

REFERENCES

[1] 1. M. Pritt and G. Chern, "Satellite Image Classification with Deep Learning," 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, 2017, pp. 1-7.

[2] L. Zhang, Z. Chen, J. Wang and Z. Huang, "Rocket Image Classification Based on Deep Convolutional Neural Network," 2018 10th International Conference on Communications, Circuits and Systems (ICCCAS), Chengdu, China, 2018, pp. 383-386.

[3] C. Shen, C. Zhao, M. Yu and Y. Peng, "Cloud Cover Assessment in Satellite Images Via Deep Ordinal Classification," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 3509-3512.

[4] T. Postadjian, A. L. Bris, C. Mallet and H. Sahbi, "Superpixel Partitioning of Very High Resolution Satellite Images for Large-Scale Classification Perspectives with Deep Convolutional Neural Networks," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 1328-1331.

[5] Q. Liu, R. Hang, H. Song and Z. Li, "Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 1, pp. 117-126, Jan. 2018.

[6] T. Postadjian, A. L. Bris, H. Sahbi and C. Malle, "Domain Adaptation for Large Scale Classification of Very High Resolution Satellite Images with Deep Convolutional Neural Networks," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 3623-3626.

[7] P. Helber, B. Bischke, A. Dengel and D. Borth, "Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 204-207. [8] K. Cai and H. Wang, "Cloud classification of satellite image based on convolutional neural networks," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2017, pp. 874-877.

[8] A. O. B. Özdemir, B. E. Gedik and C. Y. Y. Çetin, "Hyperspectral classification using stacked autoencoders with deep learning," 2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Lausanne, 2014, pp. 1-4.

[9] M. Lavreniuk, N. Kussul and A. Novikov, "Deep Learning Crop Classification Approach Based on Sparse Coding of Time Series of Satellite Data," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 4812-4815.

[10] L. Bragilevsky and I. V. Bajić, "Deep learning for Amazon satellite image analysis," 2017 IEEE Pacific Rim.