CREDIT CARD FRAUD DETECTION USING RANDOM FOREST (CART) ALGORITHM

[1]Mr. MANAS KUMAR, [2]V. LAXMI NIVAS, [3]G. NACHIKETHA & [4]M. TEJA

[1]Assistant Professor, Department of Information Technology, CMR College of Engineering & Technology

[2, 3, 4] B-Tech, Department of Information Technology, CMR College of Engineering & Technology

Abstract:

The project is mainly focused on credit card fraud detection in real world. Over few years, Credit card fraud has become one of the most common types of fraudulent issues. The recent increase in transactions has resulted with such a huge increase in illegal transactions. The purpose is to obtain things without having to pay for them or to remove money from one account without being authorized. All credit card providing institutions must implement robust detecting fraud systems in order to minimize their losses. The fact that neither cards nor card users must be present in the transaction is among the most difficult parts of running a business. As a result, the card distributer has no way of determining whether or not the customer is purchasing the actual cardholder. The accuracy of the proposed scheme is achieved by utilizing the random forest. our proposed solution, using random forest algorithm gives the accuracy of detecting the fraud

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. Credit cards are widely used due to the popularization of ecommerce and the development of mobile intelligent devices. Credit card has made an online transaction easier and more convenient. Fraud detection is a process of monitoring the transaction behaviour of a cardholder in order to detect whether an incoming transaction is done by the cardholder or others. Credit card fraud detection is a relevant problem that draws the attention onal intelligence communities, where a large number of automatic solutions have been proposed. In a real-world FDS, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Classifiers are typically employed to analyze all the authorized transactions and alert the most suspicious ones. Alerts are then inspected by professional investigators that contact the cardholders to determine the true nature (either genuine or fraudulent) of each alerted transaction. By doing this, investigators f machine-learning and computatio provide a feedback to the system in the form of labelled transactions, which can be used to train or update the classifier, in order to preserve (or eventually improve) the fraud-detection performance over time. The vast majority of transactions cannot be verified by investigators for obvious time and cost constraints. These transactions remain unlabeled until customers discover and report frauds. Another important difference between what is typically done in the literature and the real-world operating conditions of Fraud-Detection System (FDS) concerns the measures used to assess the frauddetection performance. We use random forest to train the normal and fraud behaviour features. Random forest is a classification algorithm based on the votes of all base classifiers.

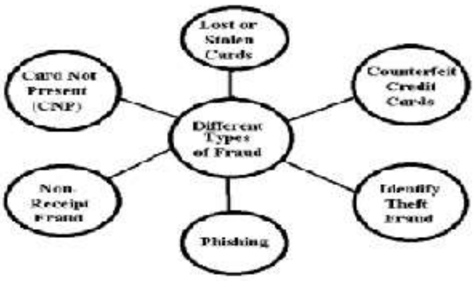

Fig 1: Different types of frauds

OBJECTIVES:

The main objective of credit card fraud detection using the Random Forest (CART) algorithm is to prevent and minimize losses due to fraudulent activities related to credit card transactions. The credit card companies can use this technique to detect fraudulent transactions in real-time or post-transaction analysis, allowing them to take necessary actions quickly to reduce the financial loss to the company and its customers. The Random Forest (CART) algorithm is a popular machine learning algorithm that can be trained on a large dataset of credit card transactions to identify patterns and anomalies that indicate fraudulent activities. The objectives of using the Random Forest (CART) algorithm for credit card fraud detection include:

- Minimizing financial losses due to fraudulent activities.
- Improving customer satisfaction by reducing the incidence of fraudulent transactions.
- Enhancing the reputation of credit card companies by providing reliable and secure payment services.
- Increasing the efficiency and accuracy of fraud detection and prevention.
- Reducing the workload and costs associated with manual fraud detection.
- Providing real-time fraud detection and prevention, enabling quick action to be taken against fraudulent activities.

IMPLEMENTATION

- Firstly, we use clustering method to divide the cardholders into different clusters/groups based on their transaction amount, i.e., high, medium and low using range partitioning.
- Using Sliding-Window method, we aggregate the transactions into respective groups, i.e., extract some features from window to find cardholder's behavioural patterns. Features like maximum amount, minimum amount of transaction, followed by the average amount in the window and even the time elapsed.
- Every time a new transaction is fed to the window the old once are removed and step-2 is processed for each group of transactions. (Algorithm for Sliding-Window based method to aggregate are referred from [1]).
- After pre-processing, we train different classifiers on each group using the cardholders' behavioral patterns in that group and extract fraud features. Even when we apply classifiers on the dataset, due to imbalance (shown in fig) in the dataset, the classifiers do not work well on the dataset. Thus, we perform SMOTE (Synthetic Minority OverSampling Technique) operation on the dataset.
- Oversampling does not provide any good results.

- Thus, there are two different ways of dealing with imbalance dataset i.e., consider Matthew Coefficient Correlation of the classifier on the original dataset or we make use of one-class classifiers.
- Finally, the classifier that is used for training the group is applied to each cardholder in that group. The classifier with the highest rating score is considered as cardholder's recent behavioral pattern.
- Once the rating score [1] is obtained, now we append a feedback system, wherein the current transaction and updated rating score are given back to the system (for further comparison) to solve the problem of concept drift.

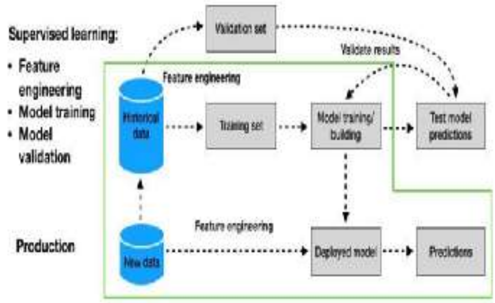

Fig 2: Flow chart

PROPOSED SYSTEM

Objective of proposed model In proposed System, we are applying random forest algorithm for classification of the credit card dataset. Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has an advantage 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 individual 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. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting. Random Forest Algorithm is used to detect the accuracy of the fraud in the transaction. Random choice forests are another name for random forests. These are a categorization, regression, and other tasks that use an ensemble learning strategy that involves teaching a greater number of decision trees and then determining the norm of the classifications (categorization) or the overall prediction (regression) of each tree. Random Forest is a supervised classification technique that uses ensemble learning. Ensemble model is a type of machine learning in which multiple versions of a same algorithm are combined to create a far more effective predictive model.

Algorithm Used for Proposed Random Forest Algorithm: Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance. Python SKLEARN inbuilt contains support for CART with all decision trees and random forest classifier. Random forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity, and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset

steps to use the Random Forest algorithm for credit card fraud detection:

1.Data Collection: Collect the transaction data from credit card companies or other sources. The data should include information such as transaction amount, transaction time, transaction location, and other relevant details.

2.Data Preprocessing: The collected data needs to be cleaned and preprocessed before applying the algorithm. This involves handling missing data, outliers, and transforming categorical variables into numerical variables.

3. Feature Selection: Select the most relevant features that contribute to the classification of fraudulent transactions. This step can be done by using statistical tests, correlation analysis, or domain knowledge.

4. Train/Test Split: Split the data into training and testing sets. The training set is used to train the Random Forest model, while the testing set is used to evaluate the model's performance.

5. Model Training: Train the Random Forest model on the training set. This involves specifying the number of trees in the forest and other hyperparameters. The model will learn to classify transactions as either fraudulent or legitimate based on the selected features.

6. Model Evaluation: Evaluate the performance of the Random Forest model on the testing set. This involves computing metrics such as accuracy, precision, recall, F1 score, and ROC curve.

7. Hyperparameter Tuning: Fine-tune the hyperparameters of the Random Forest algorithm to improve its performance. This can be done by using techniques such as grid search or randomized search.

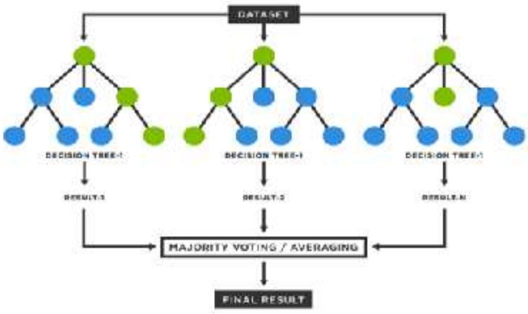

Fig 3: Random forest

RESULTS AND DISCUSSION

Comparison Of Existing Solutions In order to compare various techniques we calculate the true positive, true negative, false positive and false negative generated by a system or an algorithm and use these in quantitative measurements to evaluate and compare performance of different systems. True Positive (TP) is number of transactions that were fraudulent and were also classified as fraudulent by the system. True Negative (TN) is number of transactions that were legitimate and were also classified as legitimate. False Positive (FP) is number of transactions that were legitimate but were wrongly classified as fraudulent transactions. False Negative (FN) is number of transactions that were fraudulent but were wrongly classified as legitimate transactions by the system.

Data Collection And Performance Metrics

Using above 'CreditCardFraud.csv' file we will train Random Forest algorithm and then we will upload test data file and this test data will be applied on Random Forest train model to predict whether test data contains normal or fraud transaction signatures. When we upload test data then it will contains only transaction data no class label will be there application will predict and give the result. See below test data file

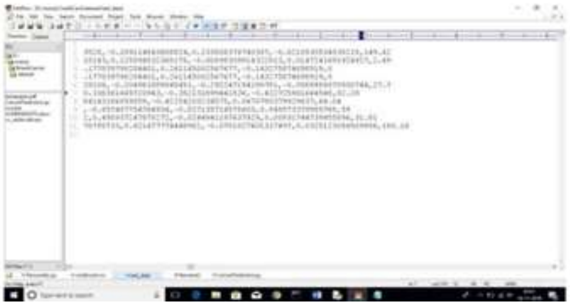

Fig 4: Data set

CONCLUSION :

The Random Forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more preprocessing techniques would also help. The SVM algorithm still suffers from the imbalanced dataset problem and requires more preprocessing to give better results at the results shown by SVM is great but it could have been better if more preprocessing have been done on the data. Random Forest Algorithm in credit card fraud detection system and the final optimization results indicates the optimal accuracy for Random Forest Algorithm is 98.6%. Although random forest obtains good results on given data set, there are still some problems such as imbalanced data. Our future work will focus on solving these problems.

FUTURE SCOPE:

It is evident from the above review that several machine learning algorithms are used to detect fraud, but the findings are not satisfactory. As a result, we'd like to use deep learning algorithms to reliably detect credit card fraud.

REFERENCES:

[1] 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.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[3] D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, "Association rules applied to credit card fraud detection," Expert Syst. Appl., vol. 36, no. 2, pp. 3630– 3640, 2009.

[4] M. Krivko, "A hybrid model for plastic card fraud detection systems," Expert Syst. Appl., vol. 37, no. 8, pp. 6070–6076, 2010.

[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Syst., vol. 50, no. 3, pp. 602–613, 2011.

[6] R. Elwell and R. Polikar, "Incremental learning of concept drift in non stationary environments," Trans. Neural Netw., vol. 22, no. 10, pp. 1517–1531, 2011.

[7] S. Jha, M. Guillen, and J. C. Westland, "Employing transaction aggregation strategy to detect credit card fraud," Expert Syst. Appl., vol. 39, no. 16, pp. 12650–12657, 2012.

[8] C. Alippi, G. Boracchi, and M. Roveri, "Just-in-time classifiers for recurrent concepts," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 4, pp. 620–634, Apr. 2013.

[9] M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero, BankSealer: A Decision Support System for Online Banking Fraud Analysis and Investigation, Berlin, Germany: Springer, 2014, pp. 380–394.

[10] C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Detecting credit card fraud using periodic features,"

in Proc. 14th Int. Conf. Mach. Learn. Appl., Dec. 2015, pp. 208–213.

[11] Reddy, b. V. R., dasari, n., & venkateswararao, k. (2021). A steganography system with gausian markov random fields and error detection codes.

[12] Revathy, G., Gurumoorthi, E., Sasikala, C., & Latha, T. M. (2023, June). Training superbot with learning automata and multi kernel SVM. In AIP Conference Proceedings (Vol. 2782, No. 1). AIP Publishing.

[13] Vinay, R., Soujanya, K. L. S., & Singh, P. (2019). Disease prediction by using deep learning based on patient treatment history. Int. J. Recent Technol. Eng, 7(6), 745-754.

[14] Challa, M. L., Soujanya, K. L. S., & Amulya, C. D. (2020). Remote monitoring and maintenance of patients via IoT healthcare security and interoperability approach. In Cybernetics, Cognition and Machine Learning Applications: Proceedings of ICCCMLA 2019 (pp. 235-245). Springer Singapore.

[15] Chandramouli, B., Vijayaprabhu, A., Arun Prasad, D., Kathiravan, K., Udhayaraj, N., Vijayasanthi, M., 2022, Design of single switch-boosted voltage current suppressor converter for uninterrupted power supply using green resources integration, Electrical Engineering and Electromechanics, 10.20998/2074-272X.2022.5.05

[16] Wang, J., Wei, K., Ansari, M.D., Al. Ansari, M.S., Verma, A., 2022, Photovoltaic Power Generation Systems and Applications Using Particle Swarm optimization Algorithms, Electrica, 10.5152/electrica.2022.22086

[17] Prasanna Moorthy, V., Siva Subramanian, S., Tamilselvan, V., Muthubalaji, S., Rajesh, P., Shajin, F.H., 2022, A hybrid technique based energy management in hybrid electric vehicle system, International Journal of Energy Research, 10.1002/er.8248

[18] Nayak, S.C., 2022, Bitcoin closing price movement prediction with optimal functional link neural networks, Evolutionary Intelligence, 10.1007/s12065-021-00592-z

[19] Rajashekhar, K., Vinod, G., Mahesh Kumar, K., Naik, J.L., 2022, Impact of erbium (Er) doping on the structural and magnetic properties of Ni-Cu (Ni0.1Cu0.9Fe2O4) nanoferrites, Journal of Magnetism and Magnetic Materials, 10.1016/j.jmmm.2022.169323