Identifying Unauthorized Transactions On Credit Cards By Using Machine Learning Methodologies

RUBEENA RAB
M.Tech Student, Dept of CSE, Malla Reddy College of Engineering and Technology, Hyderabad, T.S, India

G.RAVI
Associate Professor, Dept of CSE, Malla Reddy College of Engineering and Technology, Hyderabad, T.S, India

Dr. M. SAMBASIVUDU
Associate Professor, Dept of CSE, Malla Reddy College of Engineering and Technology, Hyderabad, T.S, India

Abstract: **It is essential for organizations that issue credit cards to be able to recognize fraudulent credit card transactions. This will prevent consumers from being charged for products that they did not buy with their credit card. The purpose of this project is to demonstrate the modelling of a data set via use of machine learning for the detection of credit card fraud. The problem of detecting fraudulent use of credit cards requires modelling previously completed credit card transactions using the information from those that were determined to be fraudulent. After that, this model is put to use to determine whether or not a new transaction constitutes fraudulent activity. Our goal is to appropriately handle misclassified categories by reducing the number of false Negative cases. During this stage of the process, our primary focuses have been on the analysis and preprocessing of data sets, as well as the application of multiple anomaly detection algorithms these algorithms include the local outlier factor and the isolation forest algorithm. We have used IEEE_CIS Fraud dataset, provided by the kaggle .we applied feature extraction technique to reduce the dimensionality of large dataset by extracting only those principle components with highest variance. Given the class imbalance ratio, we measured the accuracy using the Area Under the Precision-Recall Curve (AUPRC) which gives better results than any other previously used models.**

Keywords: **Automated Fraud Detection; Applications Of Machine Learning; Data Science; And Random Forest Algorithm;**

I. INTRODUCTION:

When a credit card is used for a purchase that was not authorized by the cardholder, this is considered "fraud" on the credit card. In order to put a stop to this abuse, the required safeguards should be applied, and the behaviour of such fraudulent activities may be analyzed in order to lessen the damage and protect against future occurrences. To phrase it another way, credit card fraud takes place when an unauthorized party makes purchases using another person's credit card while the cardholder and the card issuer are blissfully unaware of the transaction. A crucial component of fraud detection is the monitoring of user activities in order to recognize fraudulent, invasive, or defaulting behaviour [1]. It is essential that organizations like machine learning and data science pay attention to this problem since it may be able to be handled automatically by those two fields. Due to the various defining elements of this topic, such as the inequality across classes, it is challenging from a pedagogical standpoint. There are far more transactions that are legal than there are that are fraudulent. In addition to this, the statistical characteristics of the transaction patterns have a tendency to change over the course of time. In spite of the fact that these are challenges, there are still additional complications involved in putting a fraud detection system into practice. In the actual world, automatic tools do a thorough review of the never-ending stream of payment requests in order to determine which ones should be granted approval. The experts investigate these complaints by getting in touch with the cardholders to determine whether or not the transaction was legitimate. Since it has been educated and improved in response to the investigators' feedback, the automated system is becoming better at its job of identifying fraudulent activity as time goes on. The columns labelled "Time" and "Amount" will be the first ones to have their scales modified. The scale used for the other columns should be used for the time and money columns as well. In addition, we need to construct a subsample of the data frame so that our algorithms may analyze a balanced distribution of fraudulent and non-fraudulent occurrences. This distribution will be provided by the subsample. The decision tree is an inferior method to the random forest, which is preferable due to the fact that the random forest protects users from wrongly over fitting their data. The decision tree is trained using a random subset of the training data. After that, the tree branches out depending on features that are picked at random from the whole collection of features at each node. Since every tree in the random forest is trained independently from the others, the training process is highly efficient, even when applied to large data sets that include a great number of data instances and features. The random forest approach prevents over fitting of the data and allows for a generalization error to be predicted with a realistic level of accuracy.

II. PROBLEM STATEMENT:

In fraudulent activities involving mobile payments, it is common for sensitive financial information like credit card numbers or personal identification numbers to be stolen and then used to make purchases or withdrawals without permission. Mobile payment fraud is getting worse because more and more people have smart phones and use services that make it easy to convert digital currencies [2]. In order to find mobile payment fraud in the real world, you need a very precise method, since financial fraud costs money. So, we came up with a way to find fraudulent mobile payments and process a lot of financial data that is based on machine learning and can be either supervised or unsupervised. Also, thanks to our technology, the sample selection and feature selection processes can be done much faster. Obtaining a high level of accuracy in the detection of mobile payments while simultaneously processing enormous volumes of transaction data this technique places its main emphasis on the variables representing the constants and the categories. One of the most widely used approaches to machine learning is called Adaboost, and it was first developed for use in binary classification projects. The method is often employed in order to increase the effectiveness of the decision tree. This is a main application, much like the one using regression classification. In order to differentiate between fraudulent and genuine financial transactions, the Adaboost algorithm is used. After doing extensive investigation, they came to the realization that Adaboost and logistic regression both provide the most trustworthy outcomes. Because the accuracy of each algorithm is the same, determining which one is better requires looking at the amount of time it takes to run each of them [3]. After taking the detection time into consideration, they came to the conclusion that the Adaboost algorithm was successful in identifying fraudulent credit card activity. On the basis of the values of a number of different characteristics of consumer behaviour, a distance between the observed value of an attribute and its present value has been computed. For medium-sized online transactions, non-traditional approaches, such as a hybrid data mining/complex network classification algorithm, have been shown to be successful. This is because the technique is based on a network reconstruction approach, which enables the creation of representations of the divergence of one instance from a reference group. This is why this is the case. Moreover, there have been efforts made to address the problem of development from a novel vantage point. In the case that a transaction is carried out with fraudulent intent, efforts have been made to improve the alert feedback interaction [4]. In the case that a transaction was found to be fraudulent, the genuine system would be contacted, and a rejection of the one that was currently being processed would be relayed via feedback. An innovative approach to preventing and detecting fraudulent activity was provided by the artificial genetic algorithm, which took a unique slant on the problem.

III. PROPOSED METHODOLOGIES:

We suggest using a machine learning model to stop people from using stolen credit cards to buy things online. It is simply not possible to manually analyze fraudulent transactions because of the massive amount of data involved and the intricacy of the data. On the other hand, if machine learning is given access to enough and informative data, this objective may be accomplished. We are going to put this notion to the test as part of the investigation [5]. Differentiating between legitimate and fraudulent credit card transactions by the use of a supervised learning method, such as a random forest, will assist us in increasing public awareness of the fraud without putting us in danger of losing any money in the process. Researchers have examined a wide variety of tactics for identifying fraudulent behaviour in credit card transactions. Some of these strategies include methods for constructing models utilizing artificial intelligence (AI), data mining, fuzzy logic, and machine learning. Despite the ubiquitous usage of credit cards, identifying fraudulent activity on these cards continues to be a difficult and essential problem. The adoption of machine learning as a preventative measure against credit card fraud is recommended by our team. Because new techniques for artificial intelligence have been developed. It has become abundantly clear that machine learning is not the solution to the problem of fraud prevention [6]. Throughout the course of an online transaction, large volumes of data are sent back and forth, and the final outcome might be genuine or fraudulent. Online businesses have a reliable method for detecting fraudulent transactions thanks to chargeback's. The fraudulent imitation datasets are used to construct various features. In this category are included considerations such as the age of the customer's account and the amount on the account, in addition to the nation in which the credit card was issued. There is a very broad range of possibilities, and each one contributes, to varying degrees, to the possibility that there will be fraud. Since the artificial intelligence of the machine is driven by the training set and not by a fraud analyst, the degree to which each characteristic contributes to the fraud score is a product of the machine itself. To emphasize this point, if it can be shown that card fraud is widespread, then every transaction that is completed using a credit card will be subjected to a greater fraud weight. If, on the other hand, this were to decrease, the amount of

contribution would stay the same. Without the need for explicit programming or human inspection, these models are able to learn on their own. Machine learning makes use of classification and regression strategies for the purpose of identifying fraudulent credit card activity. We employ supervised learning methods like the random forest algorithm to assess whether or not a certain fraudulent credit card transaction happened online or offline. Random Forest employs an algorithm that is more complicated than the decision tree that is traditionally used. In terms of both accuracy and efficiency, the random forest method exceeds its competition in the field of machine learning. The random forest algorithm makes an effort to address the aforementioned correlation issue by picking an increasingly minute fraction of the feature space with each subsequent split. Its primary objective is to provide a stopping condition for node splits so that the trees may be trimmed and made to have a greater bearing on the aesthetics of the room. As technological advancements continue, it will be more difficult to keep up with the modelling and pattern of illegal activities. Due to developments in machine learning, artificial intelligence, and a number of other significant areas of information technology, it is now possible to automate this process and save some of the substantial amount of labour that is put into determining whether or not a transaction involves fraudulent use of a credit card.

IV. ENHANCED SYSTEM:

This evaluation is based on a database of reviews of different items written by consumers. These reviews were taken from their credit card statements. At this point, you are going to take the vast quantity of data at your disposal and reduce it to a more reasonable level. When trying to solve ML problems, it helps to have a lot of data (examples or observations) for which you know what the desired result is. Data that has been labelled allows you to anticipate the result. The term "data cleaning" is used to describe the steps taken to rectify and standardize information prior to its use. It is important to have a standard format for input files. The.csv file extension is strongly recommended. To ensure the success of any data science Endeavour, it is essential to first cleanse the data. A few of these instances include filling in missing information and streamlining classification. Much of a data scientist's time is spent on the tedious but essential task of cleansing data. It's a strategy for getting meaningful insights out of large datasets by drilling down into more manageable subsets. Data visualization is the act of presenting information in a form that is readily consumable by the human eye, and data scientists convey stories with the visualizations they generate. When it comes to working with data and generating high-quality results, Tableau is the best software available. In order to extract features from the examined data for use in testing and training, an analysis of the data's behaviour and pattern is required. Our models are then fine-tuned through the classifier technique. The classify module from the Natural Language Toolkit package in Python is used. It makes use of the acquired, labelled dataset. We will use the remaining labelled data we have to test the models. Machine learning strategies were used to classify the data ahead of time. Classifiers based on a random forest algorithm were chosen. This strategy is often employed for the purpose of classifying texts. Improving a model requires first assessing how well it already functions. This helps in deciding which model is the greatest match for our data and how long-lasting the selected model will be. Evaluation of model performance using the same data used for training is seldom accepted in data science since it is easy to construct overly optimistic and over fit models. Over fitting may be tested via performance evaluation techniques like holdout and cross-validations. Visual representations of the results will be provided. Information that has already been organized may be represented in graphs. An accurate model is one that makes a large proportion of right predictions based on test data. The solution may be found by dividing the number of correct guesses by the total number of predictions. The random forest method nevertheless produces accurate results, despite the fact that the data may be unbalanced or under scaled. So, by analysing the provided information using the random forest approach and the decision tree methodology, we were able to derive a precise percentage of fraud detection. Confusion matrices provide a concise summary of a classifier's performance on test data when the true values are known. You can easily categories your data and see the outcomes of your algorithms. So, most performance indicators may be calculated by learning more about the classification model's error rates and the factors that contribute to those rates.

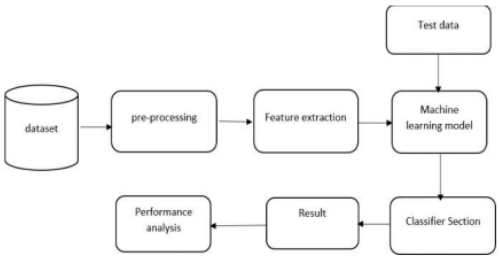

Fig 1: System Architecture

V. CONCLUSION:

Making use of a credit card that has been stolen is unequivocally dishonest. This article examines recent advancements in this industry and outlines the most common forms of fraud, as well as ways in which these types of fraud may be discovered. This article includes a discussion of how machine

learning may be used to improve the accuracy of this process. In addition to the method, pseudo code, description of its implementation, and experimental findings for detecting fraud, this article also includes a discussion of how machine learning may be used. The proposed module may be used on a larger dataset and produce findings that are more reliable than those produced using the modules that are currently being used. Even though the performance of the Random Forest algorithm has improved with more training data, it is still behind the competition when it comes to the speed with which it performs tests and is used in practice. It's possible that using pre-processing processes more often may be beneficial as well. In upcoming projects, one of our goals is to develop anti-credit card fraud software that makes use of cutting-edge technology such as machine learning, artificial intelligence, and deep learning. This will allow us to better identify fraudulent activity and take appropriate action. The dataset offers room for further expansion and improvement. It has been shown in the past that using a more extensive dataset results in more precise algorithms. So, increasing the quantity of data will unquestionably increase the model's capacity to detect fraudulent activity while at the same time reducing the number of false positives. Having said that, official support from the banks is required.

REFERENCES:

[1] "Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[2] A. Charleonnan, "Credit card fraud detection using RUS and MRN algorithms," 2016 Management and Innovation Technology International Conference (MITicon), Bang-San, 2016.

[3] M. Kavitha and M. Suriakala, "Hybrid Multi-Level Credit Card Fraud Detection System by Bagging Multiple Boosted Trees (BMBT)," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, 2017.

[4] M. Kavitha and M. Suriakala, "Real time credit card fraud detection on huge imbalanced data using meta-classifiers," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017.

[5] M. F. Zeager, A. Sridhar, N. Fogal, S. Adams, D. E. Brown and P. A. Beling, "Adversarial learning in credit card fraud detection," 2017 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2017.

[6] Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, Francesco Palmieri, Using generative adversarial networks for improving classification effectiveness in credit card fraud detection, Information Sciences, Volume 479, 2019.

[7] F. Ghobadi and M. Rohani, "Cost sensitive modeling of credit card fraud using neural network strategy," 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS), Tehran, 2016.

[8] A. Agrawal, S. Kumar and A. K. Mishra, "Implementation of Novel Approach for Credit Card Fraud Detection," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 1-4.

