International Journal of Recent Advances in Multidisciplinary Topics
Volume 2, Issue 4, April 2021
https://www.ijramt.com | ISSN (Online): 2582-7839

Credit Card Fraud Detection Using Machine Learning

Sanisa Saiju[1*], S. Akshaya Jyothy[2], Christeena Sebastian[3], Liss Mathew[4], Tintu Sabu[5]
[1,2,3,4,5]Department of Computer Application, Saintgits College of Engineering, Kottayam, India

***Abstract*: Machine learning and AI techniques are becoming utilized in conjunction with processing to unravel a superfluity of real world issues. These techniques have proved to be extremely effective, yielding most accuracy with nominal financial investment and additionally saving immense amounts of it slow. To feature to their annual financial gain, nowadays, folks have started look stock investments as a moneymaking possibility. With knowledgeable steerage and intelligent designing, it wills virtually double the annual revenue through stock returns. That said, stock investment still remains a risky proposition for the inexperienced. Usurious wages of the investment consultants as well as a general content regarding the monetary matters among the overall public, deters several from commerce in stocks. The worry of losses additionally acts of stocks as a deterrent to many. These facts propelled to harness the ability of machine learning to predict the movement victimization sentiment analysis on the tweets collected victimization the Twitter API and additionally the closing values of varied stocks, seeked to create a system that forecasts the stock value movement of varied corporations. Such a prediction would greatly facilitate a potential stock capitalist in taking wise to choices which could directly contribute to his profits.**

***Keywords*: Applications of Machine Learning, Automated fraud detection, Credit card fraud, Data science, Isolation forest algorithm, Local outlier factor.**

1. Introduction

Credit card fraud may be a major problem that involves payment card like master card as illegal source of funds in transactions. Fraud is illegal thanks to obtain goods and funds. The goal of such illegal transaction could be to urge products without paying or gain an unauthorized fund from an account. Identifying such fraud may be a troublesome and should risk the business and business organizations. In the real world FDS investigator are not able to check all transactions. Here the Fraud Detection System monitors all the approved transactions and alerts the foremost suspicious one. A customer provides FDS with feedback if the transaction was authorized or fraudulent. Verifying all the alerts everyday could even be a time consuming and dear process. The remainder of the transactions remains unchecked until customer identifies them and reports them as a fraud. Also the techniques used for fraud and therefore the cardholder spending behavior changes over time. This change in master card transaction is named as concept drift. Most of the time it's difficult to spot the master card fraud. Machine Learning is taken into account together of the foremost successful technique for fraud identification. It uses classification and regression approach for recognizing fraud in master card. The machine learning algorithms are divided into two types, supervised and unsupervised learning algorithm. Supervised learning algorithm uses labeled transactions for training the classifier whereas unsupervised learning algorithm uses coeval's analysis that groups customers consistent with their profile and identifies fraud based on customers spending behavior. Machine learning algorithms are employed to analyze all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to verify if the transaction was genuine or fraudulent. Detection methods are developed to defend criminals in adapting to their fraudulent strategies. Many learning algorithms are presented for fraud detection in master card which incorporates neural networks, Logistic Regression, decision tree, Naive Bayes, Support Vector Machines, K-Nearest Neighbors and Random Forest. This project will be introducing performance of above algorithms supported their ability to classify whether the transaction was authorized or fraudulent then compares them. The comparison is formed using performance measure accuracy, specificity and precision. It shows that Random Forest algorithm showed better accuracy and precision than other techniques.

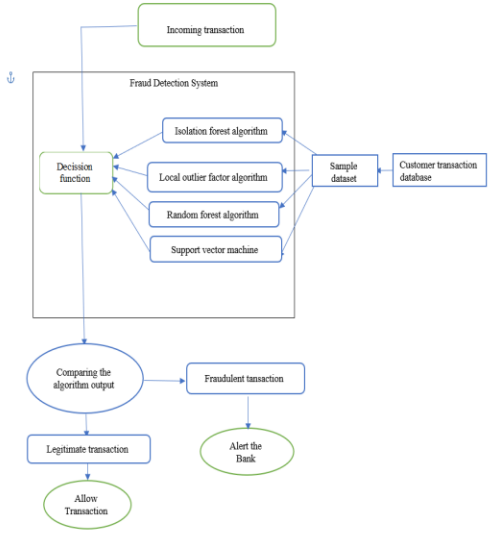

Fig. 1. Flowchart

*Corresponding author: sanisasaiju1998@gmail.com

2. Literature Review

Fraud acts because the unlawful or criminal deception supposed to lead to money or personal profit. It's a deliberate act that's against the law, rule or policy with AN aim to realize unauthorized money profit. Varied literatures relating anomaly or fraud detection during this domain are revealed already and square measure obtainable for public usage. A comprehensive survey conducted by Clifton Phua and his associates have disclosed that techniques utilized during this domain embrace data processing applications, machine-driven fraud detection, adversarial detection. In another paper, Suman, analysis Scholar, GJUS&T at Hisar HCE bestowed techniques like supervised and unattended Learning for master card fraud detection. Although these ways and algorithms fetched an sudden success in some areas, they didn't offer a permanent and consistent answer to fraud detection. an analogous analysis domain was bestowed by Wen-Fang YU and atomic number 11 Wang wherever they used Outlier mining, Outlier detection mining and Distance total algorithms to accurately predict fallacious dealing in an emulation experiment of master card dealing information set of 1 bound depository financial institution. Outlier mining could be a field of knowledge mining that is essentially utilized in financial and web fields. It deals with detection objects that square measure detached from the most system i.e. the transactions that aren't real. they need taken attributes of customer's behavior and supported worth the worth of these attributes they've calculated that distance between the discovered price of that attribute and its planned value. Unconventional techniques like hybrid information mining/complex network classification rule is ready to understand banned instances in an actual card dealing information set, supported network reconstruction rule that permits making representations of the deviation of 1 instance from a reference cluster have evidenced economical generally on medium sized on-line dealing. There have conjointly been efforts to progress from a very new side. Tries are created to enhance the alert feedback interaction just in case of fallacious dealing. Just in case of fallacious dealing, the licensed system would be alerted and a feedback would be sent to deny the continuing dealing. Artificial Genetic rule, one among the approaches that shed new lightweight during this domain, countered fraud from a special direction. It evidenced correct to find out the fallacious transactions and minimizing the amount of false alerts. Although, it had been amid classification drawback with variable misclassification prices.

3. Methodology

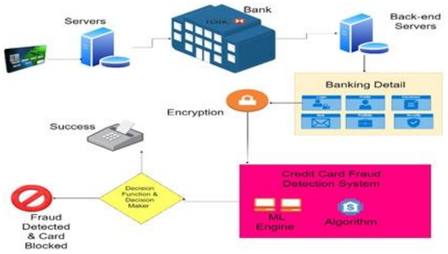
Servers
Bank
Back-end Servers
Banking Detail
Encryption
Success
Credit Card Fraud Detection System
ML Engine
Algorithm
Decision Function & Decision Maker
Fraud Detected & Card Blocked

Fig. 2. Working of credit card fraud detection system

The approach that this paper proposes is to use the latest machine learning algorithms to detect anomalous activities and these activities are called outliers. The approach that this paper proposes is to use the latest machine learning algorithms to detect anomalous activities and these activities are called outliers. When considering the real life, the architecture could be described as this, we obtained the dataset from Kaggle; it is a data analysis website which provides datasets. The dataset has 31 columns out of which 28 are named as v1-v28 this is done in order to protect sensitive data. The other three columns are Time, Amount and Class. Time showed the gap between the first transaction and the following transaction or the time limit or variation between the two. Amount is the amount of money that is used during the transaction. Class has two values o and 1, 0 represents a valid transaction and 1 represents a fraudulent one. We have used various graphs in order to comprehend the data and view the data in an graphical way.

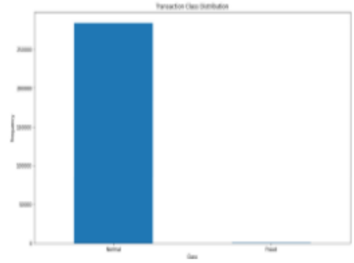

Fig. 3. Count of fraudulent vs. non-fraudulent transactions

When we take abundant a way lay far clearer look at the group action we are able to see that the quantity used for fraudulent transactions are much less than that of real transactions. Once this analysis, we tend to plot a heat map to induce a colored illustration of the info and to check the correlation between out predicting variables and therefore the category variable. This heatmap is shown below:

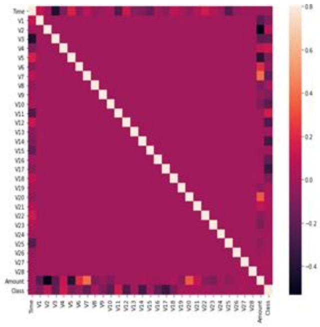

Fig. 4. Correlation between predicting and class variable

The dataset is then formatted and processed. The time and amount column are standardized whereas the Class column is removed to ensure fairness of evaluation when performed. The data is processed by using a set of algorithms from different modules. The data is then fitted into a model and the following outlier detection modules are applied on it:

- Local Outlier Factor
- Isolation Forest Algorithm

- Random Forest Algorithm
- Support Vector machine

These algorithms are a part of sklearn module. The ensemble module in the sklearn package includes ensemble based methods and functions that would be used for classification, regression and outlier detection. It is a free and open-source Python library that is built using NumPy, SciPy and matplotlib modules and those modules provide a lot of simple and efficient tools which could be used for data analysis and also for machine learning. We've used the Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests. This program can also be executed using Google Collab platform on the cloud and it supports all python notebook files. Detailed explanations about the modules and algorithms are given as follows:

1) Local Outlier Factor

It is associate unattended Outlier Detection rule. 'Local Outlier Factor' refers to the anomaly score of every sample. It measures the native deviation of the sample knowledge with relation to its neighbours.' k-nearest neighbors', whose distance is employed to estimate the native knowledge. The native values of a sample thereto of its neighbours, one will determine samples that are well under their neighbours. These values are quite amorous and that they are thought of as outliers. Because the dataset is extraordinarily massive, we have a tendency to used solely a fraction of it in out tests to cutback interval

2) Isolation Forest Algorithm

It is one among the latest techniques to find anomalies. The rule is works on the very fact that anomalies are knowledge points that are few however totally different. And as results of these properties, anomalies are prone to a mechanism known as isolation. Isolation Forest rule works like this, it isolates observations by indiscriminately choosing a feature and so indiscriminately choosing a split worth which would be in between the most and minimum values of the chosen feature. The logic argument would go like this; uninflected anomaly observations is less complicated as a result of solely a number of conditions are required to separate the fraudulent cases from the conventional observations. However, we might need a lot of conditions to isolate traditional conditions.so, we will calculate the anomaly score because the range of conditions needed to separate a given observation. The means that the rule constructs the associate separation is by 1st making isolation trees, and then, the score is calculated because the path length to isolate the observation.

3) Random Forest Algorithm

Random forest may be a supervised machine learning rule that's supported the conception of ensemble learning. The random forest rule works in a very similar thanks to that of isolation forest and uses multiple algorithms i.e. multiple call trees, leading to a forest of trees, therefore the name "Random Forest". The random forest rule is usually used for each regression and classification tasks.

4. Implementation

This idea is troublesome to implement in world as a result of it needs the cooperation from banks that will not be willing to share data because of their market competition, and additionally because of legal reasons and protection of data of their users. Therefore, we have a tendency to search some reference papers that followed similar approaches and gathered results. As declared in one in all these reference papers: "This technique was applied to a full set of information equipped by a German bank in 2006. For banking there'll be several problems like confidentiality reasons. Solely an outline of the results obtained is bestowed below. When applying this method, the extent one list encompasses a few of cases however with a high likelihood of being fraudsters. All people mentioned throughout this list had their cards closed to avoid any risk because of their speculative profile. The condition is a lot of advanced for the opposite list. The extent a pair of list continues to be restricted to be checked on a case by case basis. Credit and assortment officers' thought-about that0.5 the cases during this list can be thought-about as suspicious dishonest behavior. For the last list and thus the most important, the work is equitably significant. But a 3rd of them area unit suspicious. so as to maximize the time potency and thus the overhead charges, a chance is to include a replacement component inside the query; this component is commonly the 5 1st digits of the phone numbers, the e-mail address, and thus the watchword, as associate degree example, those new queries is commonly applied to the extent a pair of list and level three lists."

5. Results

We have found the total number of false positives it is then compared with the actual values. The result obtained from this is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster testing is 10% of the entire dataset. We have done this to reduce the processing time as the dataset is huge we would require a huge amount of time to process all the four algorithms. The results along with the classification report for each algorithm is given in the screenshot and the output as follows, class 0 represents that the transaction was determined to be valid and 1 represents that it was determined as a fraud transaction. The result that we have obtained is then matched against the class values to check for false positives. Random forest classifier detected 4 errors versus Isolation forest detected 77 detected errors versus local outlier factor detecting 97 errors versus support vector Machines detecting 8516 errors Random forest classifier has a 99.92 % more accuracy than Isolation forest which has an accuracy of 99.72 % versus Local outlier factor which has an accuracy of 99.65 % and SVM which has an accuracy of 70.09 % when comparing error precision and Recall for 4 models , the Random forest classifier performed much better than the isolation forest because it detects fraud cases at 67% ,isolation detection rate is 22 % and LOF detection rate is just 2 % and SVM of 0 %. So the overall Random forest method performed much better in determining the fraudulent cases. We can also improve this accuracy at computational cost by increasing the sample size or by using a machine learning algorithm. We can use complex error detection models to achieve better accuracy in determining more fraud cases.

6. Conclusion

Credit card fraud could be a doubt and criminal act. This text has listed out the strategies to search out dishonorable transactions together with their detection strategies and reviewed recent findings during this field. This paper has conjointly explained intimately, however machine learning are often applied to induce higher leads to fraud detection and that formula works higher during this field with clarification and details of the method together with the result. once the formula reaches 99% accuracy, its accuracy can stay at twenty eighth for one tenth of the dataset. However, once the entire dataset is gift given the formula, the accuracy will increase to thirty third. This high proportion of accuracy is to be expected thanks to the large imbalance between range the amount the quantity of valid and number of real transactions. Since the whole dataset contains solely 2 days of group action records, solely some of its knowledge are obtainable if the project is to be used commercially. as a result of it's supported a machine learning formula, the program can solely increase its potency over time because it provides additional knowledge.

7. Future Enhancements

While we cannot achieve 100% accuracy in fraud detection, we have created a system that can get very close to that goal with enough time and data. As with any such project, there is little room for improvement here. The nature of this project allows multiple algorithms to be combined into modules, and combining their results to increase the accuracy of the final result. This model can be further improved by adding more algorithms. However, the output of these algorithms must be in the same format as the others. Once that is done, it is easy to add modules as coded. This gives the project a lot of modularity and variety. More space for improvement can be found in the dataset. As previously demonstrated, the accuracy of algorithms increases as the size of the dataset increases. Therefore, more data will definitely make the model more accurate in detecting scams and reduce the number of false positives. However, this requires the full support of the banks themselves.

References

[1] John Richard, D. Kho, Larry, A. Vea, "Credit Card Fraud Detection Supported Dealing Behaviour," IEEE Region 10 Conference (TENCON), Malaysia, November 2017.

[2] Clifton Phua, Vincent Lee, Kate Smith and Ross Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," School of Business Systems, School of Knowledge Technology Monish University.

[3] Dal Pozzolo, Andrea Boracchi, Giacomo Caelen, Olivier Alippi, Cesare Bontempi, and Gianluca, "Credit Card Fraud Detection: A practical Modelling and a Completely Unique Learning Strategy," IEEE Dealing of Neural Networks and Learning System vol. 29, no.8, pp. 1-14, August 2017.

[4] Wen-Fang YU and Na Wang, "Credit Card Fraud Detection Model supported Distance Sum", in *International Joint Conference on Artificial Intelligence* in 2009.

[5] Ishu Trivedi, Monika, Mrigya, Mridushi, "Credit Card Fraud Detection," in *International Journal of Advanced Research in Computer and Communication Engineering* vol. 5, no. 1, January 2016.

[6] Massimiliano Zanin, Miguel Romance, Regino Criado, and Santiago Moral, "Master Card Fraud Detection through Parenclitic Network Analysis," Hindawi Complexity, 2018.