Original Research Paper

Volume - 13 | Issue - 05 | May - 2023 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Machine Learning

CREDIT CARD FRAUD DETECTION SYSTEM PROJECT

Khyati Sharma Final Year Student, Department of Computer Science and Engineering, Dronacharhya Group of Institution, Greater Noida, UP, India

Dakshita Verma* Final Year Student, Department of Computer Science and Engineering, Dronacharhya Group of Institution, Greater Noida, UP, India *Corresponding Author

Prof. Priya Rani Professor, Department of Computer Science and Engineering, Dronacharhya Group of Institution, Greater Noida, UP, India

ABSTRACT This Project is focused on credit card fraud detection in real world scenarios. Nowadays credit card frauds are drastically increasing in number as compared to earlier times. Criminals are using fake identities and various technologies to trap the users and get the money out of them. Therefore, it is essential to find a solution to these types of frauds. So, in this proposed project, we have created a Web App for the detection of such types of frauds with the help of Machine Learning. We have designed a model to detect the fraud activity in credit card transactions. This system can provide most of the important features required to detect illegal and illicit transactions. As technology changes constantly, it is becoming difficult to track the behavior and pattern of criminal transactions. To come up with the solution one can make use of technologies with the increase of machine learning, artificial intelligence and other relevant fields of information technology; it becomes feasible to automate this process and to save some of the intensive amounts of labor that is put into detecting credit card fraud.

KEYWORDS :

I. INTRODUCTION

In today's world, we are on the express train to a cashless society. According to the World Payments Report, in 2016 total non-cash transactions increased by 10.1% from 2015 for a total of 482.6 billion transactions! That's huge! Also, it's expected that in future years there will be a steady growth of non-cash transactions as shown below: Now, while this might be exciting news, on the flip-side fraudulent transactions are on the rise as well. Even with EMV smart chips being implemented, we still have a very high amount of money lost from credit card fraud: This is now becoming a serious problem since most of the time, a person who has become a victim of this fraud doesn't have any idea about what has happened until the very end. Credit card generally refers to a card that is assigned to the customer (cardholder), usually allowing them to purchase goods and services within credit limit or withdraw cash in advance. A credit card provides the cardholder with an advantage of time, i.e., it provides time for their customers to repay later in a prescribed time, by carrying it to the next billing cycle. Credit card frauds are easy targets. Without any risks, a significant amount can be withdrawn without the owner's knowledge, in a short period. Fraudsters always try to make every fraudulent transaction legitimate, which makes fraud detection very challenging and difficult task to detect. With different frauds, mostly credit card frauds, often in the news for the past few years, frauds are in the top of mind for most the world's population. Credit card dataset is highly imbalanced because there will be more legitimate transaction when compared with a fraudulent one. As advancement, banks are moving to EMV cards, which are smart cards that store their data on integrated circuits rather than on magnetic stripes, have made some on-card payments safer, but still leaving card-not-present frauds on higher rates.

Even then there are chances for thieves to misuse credit cards. There are many machine learning techniques to overcome this problem.

So, in this project, what we have tried is to create a Web App for the detection of such types of frauds with the help of Machine Learning.

II. LITERATURE REVIEW

Fraud acts as 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, and adversarial detection. In another paper, Suman, Research Scholar, GJUS&T at Hisar HCE presented techniques like Supervised and Unsupervised Learning for credit card fraud detection. Even though these methods and algorithms fetched an unexpected success in some areas, they failed to provide a permanent and consistent solution to fraud detection. A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of credit card transaction data set of one certain commercial bank. Outlier mining is a field of data mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e., the transactions that aren't genuine. They have taken attributes of a customer's behavior and based on the value of those attributes they've calculated the distance between the observed value of that attribute and its predetermined value. Unconventional techniques such as hybrid data mining/complex network classification algorithm can perceive illegal instances in an actual card transaction data set, based on network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group have proved efficient typically on medium sized online transaction. There have also been efforts to progress from a completely new aspect. Attempts have been made to improve the alert feedback interaction in case of fraudulent transactions. In case of fraudulent transaction, the authorized system would be alerted, and feedback would be sent to deny the ongoing transaction. Artificial Genetic Algorithm, one of the approaches that shed new light in this domain, countered fraud from a different direction. It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

III. Data Collection

The dataset is collected from Kaggle. The dataset provides credit card transactions done by European cardholders in September 2013 and this transaction was done in two days. In this dataset, we have found that there were only 492 frauds cases out of 284,807 transactions that occurred in the last two days. The dataset is heavily skewed, with positive class representing only 0.172 percent of all transactions. 423 Proceedings of the 2nd Indian International Conference on Industrial Engineering and Operations Management Warangal, Telangana, India, August 16-18, 2022, © IEOM Society International It has only numerical input variables most of these are PCA transformed. So V1; V2; V3; to V28 is the main component obtained using PCA; the only characteristics not changed through PCA are 'Time', 'class' and 'Amount.' Another variable is 'Class,' and it has a value of 1 which indicates frauds and o indicates genuine one. So, we have a training dataset for performing the task.

IV. Model Training

Understanding the data and related constraints

Since the data for this project is very unbalanced because number of cases of Fraud transactions are very low in comparison to number of cases of Valid transactions makes the model training a bit hectic. Because if we consider "classification accuracy" as the metric for training, we won't be getting the perfect view of how much our model is learning, because the classification accuracy is derived as:

classification_accuracy = No. of correct predictions/No. of labels So, now, consider the fact that if our data have 98% of the values to be valid while only 2% to be frauds, if our model predicts all values to be valid, it will eventually achieve 98% accuracy at the end of the day, but the model will be an absolute wastage. For this reason, we use a different type of metric which will give us much more important information about what our model has learned. Actually, what we do is print the classification matrix for our model predictions and then we judge our model based on that matrix. Precision and Recall are two of the derivatives of a confusion matrix, we will consider them both though, but only Recall will come in handy for us since high Recall will ensure that no fraud value gets detected to be a valid one. Also, Precision does the vice-versa. We will need to find the best threshold where the Precision-Recall tradeoff will give us the best results.

V. Preprocessing data

Balancing data Since the data is very imbalanced, we will be using some Under sampling and Oversampling techniques. As the name suggests, under sampling is used to reduce the samples from majority class and Oversampling is used to increase the samples from minority class. This way, we can achieve some balancing of the data. Scaling features Now, even though almost all the features are dimensionally reduced using some dimensionality reduction technique, two of the features are in their original form. Time and Amount are the two features which we will be scaling in order to make our model learn features correctly. Splitting the data Now, since we needed to save some entries from the data for our testing purpose, we will now be splitting the data into two parts, namely, train and test. Miscellaneous Now, other than the above three techniques, we did some Exploratory Data Analysis [EDA] on the data, we get the idea of outliers in the data, feature importance etc. by doing this amazing part. One can find the EDA in the notebook itself.

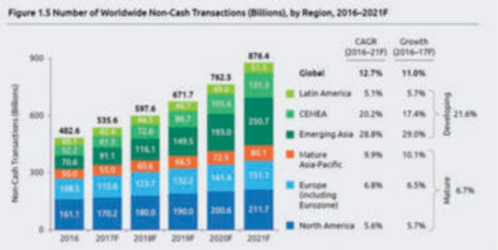
VI. Model Architecture

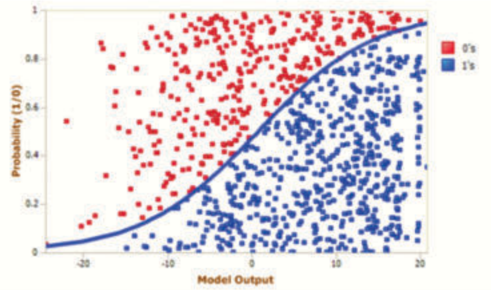
We will be using a Logistic Regression classifier for our project. A logistic Regression model is used to predict the probability of a certain class or event existing. We then decide the class from which the entry belongs by using a threshold value. This threshold value is decided by manipulating the Precision-Recall tradeoff as explained above.

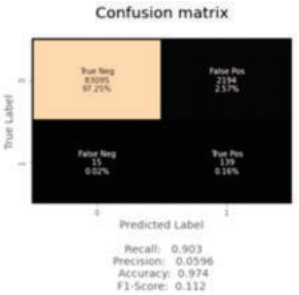
Also, while training, we used GridSearchCV to find the best possible parameters for our model so that it can squeeze the maximum amount of important information out of the data.

VII. Post Training

After a model gets trained, this is our duty to understand the results and ensure its reliability because this will be used for a generalization purpose. Let's now see the inference results we got after training by using the following confusion matrix from our training notebook:

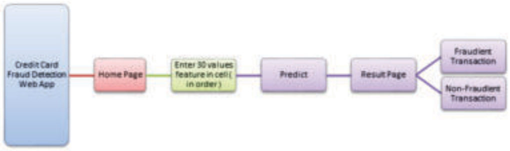






VIII. Web Application

In our web App, there is a home page & a result page. We need to do is just copy the data entry which we want to predict into the box, this box will take 30 float/integer values as input, with at least one whitespace dividing them. Now, after entering the data, we will click on the predict button. Then on the result page, result will show whether the given data is a Fraudulent transaction or not.


Credit Card Fraud Detection Web App
Home Page
Enter 30 values feature in cell (in order)
Predict
Result Page
Fraudient Transaction
Non-Fraudient Transaction

IX. Conclusion

Credit Card is a great tool to pay money easily, but as with all the other monetary payment tools, reliability is an issue here too as it is subjected to breach and other frauds. To encounter this problem, a solution is needed to identify the patterns in the transactions and identify the ones which are fraud, so that finding such transactions beforehand in future will be very easy. Machine Learning is a great tool to do this work since Machine Learning helps us in finding patterns in the data. Machine Learning can help produce great results if provided with enough amount of data. Also, with further advances in technology, Machine Learning too will advance with time, it will be easy for a person to predict if a transaction is fraud or not much more accurately with the advances.

REFERENCES

[1]. https://towardsdatascience.com/the-random-forestalgorithm-d457d499ffcd

[2]. https://www.xoriant.com/blog/productengineering/decision-trees-machine-learningalgorithm.html

[3]. Gupta, Shalini, and R. Johari. ”A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant.” International Conference on Communication Systems and Network Technologies IEEE, 2021:22-26.

[4]. Y. Gmbh and K. G. Co, “Global online payment methods: the Full year 2020,” Tech. Rep., 3 2020.

[5]. Bolton, Richard J., and J. H. David. “Unsupervised Profiling Methods for Fraud Detection.” Proc Credit Scoring and Credit Control VII (2020): 5– 7.

[6]. Drummond, C., and Holte, R. C. (2019). C4.5, class imbalance, and cost sensitivity: why under-sampling beats oversampling. Proc of the ICML Workshop on Learning from Imbalanced Datasets II, 1–8.

[7]. Quah, J. T. S., and Sriganesh, M. (2020). Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4), 1721-1732