Credit Card Fraud Detection Using Machine Learning

SUSHANT AGRAWAL
Department of Information Technology, Maharaja Agrasen Institute of Technology

Abstract- ***For clients to avoid being charged for products they did not buy, credit card issuers must be able to recognise fraudulent credit card transactions. Data Science may be used to solve issues, and coupled with machine learning, its significance cannot be understated. With the use of credit card fraud detection, this research aims to demonstrate the modelling of a data set using machine learning. The Credit Card Fraud Detection Problem includes modelling prior credit card transactions using data from those that turned out to be fraudulent. This technique then determines the validity of a new transaction. The goal here is to minimise inaccurate fraud categories while detecting 100% of the fraudulent transactions. A classic example of categorization is the detection of credit card fraud. The analysis and pre-processing of data sets, as well as the use of several anomaly detection techniques to PCA-transformed Credit Card Transaction data, have been the main points of this approach.***

I. INTRODUCTION

The illegal and unwanted use of a credit card account by someone other than the account owner is referred to as credit card fraud. The abuse may be halted by taking the required preventative measures, and it is possible to investigate the behaviour of such fraudulent activities to minimise them and avoid recurrence. Credit card fraud is, in other words, the use of someone else's credit card for personal advantage when neither the cardholder nor the entity in charge of issuing the card are aware that the card is being used.

Monitoring user populations' behaviour is a crucial part of detecting fraud since it enables the identification, detection, and prevention of unwelcome behaviours including fraud, intrusion, and defaulting.

Fields like machine learning and data analytics, where an automated solution is available, must handle this extremely important issue.

This problem is particularly challenging from a learning perspective since it exhibits several traits, such class imbalance. There are many more honest than dishonest trades. Additionally, the transaction patterns' statistical properties regularly change over time.

However, these are not the only challenges facing the creation of a practical fraud detection system. In examples drawn from the real world, automated technologies quickly examine the vast amount of payment requests to choose which transactions to approve.

Machine learning algorithms are employed to analyse all authorised transactions and identify those that appear dubious. In order to validate if the transaction was real or fraudulent, investigators who are looking into these claims contact the cardholders.

To steadily improve the accuracy of fraud detection over time, the automated system uses information from the investigators to train and update the algorithm.

To stop fraudsters from changing their deceptive strategies, techniques for detecting fraud are always being developed. Card theft, account bankruptcy, device intrusion, application fraud, counterfeit cards, telecommunication fraud, and credit card fraud both online and offline are the several types of frauds that fall under this category.

Some of the techniques used today to spot this fraud include Artificial Neural Networks, Fuzzy Logic, Genetic Algorithms, Logistic Regression, Decision Trees, Support Vector Machines, Bayesian Networks, Hidden Markov Models, and K-Nearest Neighbour.

II. LITERATURE REVIEW

Fraud is defined as an unlawful or criminal deception intended to generate financial or personal gain. In order to earn unrecognised financial gain, it is a deliberate activity to break a law, regulation, or policy.

A lot of publicly available information has previously been published in this sector on the subject of anomaly or fraud detection. According to a comprehensive analysis conducted by Clifton Phua and his colleagues, some of the techniques employed in this sector include adversarial detection, automated fraud detection, and data mining applications. In a different study, Suman—a research assistant with GJUS&T at Hisar HCE—discussed techniques for identifying credit card fraud, such as supervised and unsupervised learning. Despite the unexpected success that these methods and algorithms had in certain cases, they were unable to provide a dependable, long-term fraud detection solution.

In a study modelling credit card transaction data from a specific commercial bank, Wen-Fang YU and Na Wang employed distance sum methods, outlier mining, outlier detection mining, and outlier detection mining to accurately anticipate fraudulent transactions. The financial and internet industries make the greatest use of the field of outlier mining in data mining. It focuses on locating components that are cut off from the main system or transactions that are fraudulent. Utilizing the features of consumer behaviour and their associated values, the gap between the observed value of an attribute and its planned value was calculated.

On medium-sized online transactions, unusual techniques like hybrid data mining/complex network classification algorithms—which are based on the network reconstruction algorithm and enable the creation of representations of the deviation of one instance from a reference group—have typically been successful. These techniques can find instances of illicit activity in a real card transaction data set. There have also been endeavours to move forward from a completely other angle. There have been efforts made to improve the alert feedback interaction in the case of a fraudulent transaction.

In the event of a fraudulent transaction, feedback would be sent to the authorised system to deny the present transaction.

Artificial Genetic Algorithm was one technique that offered fresh perspective into this subject and dealt with fraud in a different approach.

It effectively detected fraudulent transactions and decreased the frequency of false alerts.

III. METHODOLOGY

The approach recommended by the study takes use of cutting-edge machine learning methods to spot outliers—unusual behaviours.

First of all, Kaggle, a website for data analysis that provides datasets, is where we received our dataset.

This dataset has 31 columns in total, 28 of which are labelled v1-v28 to protect sensitive data.

The other columns correspond to Time, Amount, and Class. Time shows the amount of time that has passed between the first and second transactions. The amount refers to the total amount of money traded. A transaction is classified as Class 0 if it is legal, and Class 1 if it is fraudulent.

After checking this dataset, we plot a histogram for each column. In order to examine if any values are missing from the dataset, it is necessary to generate a graphical representation of it. This is done to ensure that the dataset can be analysed by machine learning algorithms without the requirement for missing value imputation.

Now that the dataset has been prepared and processed. The class column has been removed, and the time and quantity columns have been harmonised, to ensure grading is fair. Data is processed by a collection of modules' algorithms.

The free and open-source Python library, which provides a variety of simple and efficient tools for data analysis and machine learning, is created by combining NumPy, SciPy, and matplotlib modules. It features a range of classification, clustering, and

regression algorithms and is designed to interface with scientific and numerical libraries. Using the Jupyter Notebook platform, we developed a Python application to demonstrate the method recommended in this paper. The Google Collab platform, which accepts any files written with Python Notebook, may also be used to execute this application on the cloud.

IV. IMPLEMENTATION

It is difficult to put this idea into practise since banks are required yet are unwilling to collaborate due to market competition, legal issues, and the need to protect consumer data. We looked for numerous reference works that employed similar methods in order to gather knowledge. A thorough application data set given by a German bank in 2006 was subjected to this method, according to one of these reference studies. Due to concerns about banking secrecy, just a summary of the findings is provided here. This technique results in a tiny number of instances on the level 1 list, but these examples are very likely to be fraudsters.

All of the individuals on this list had their cards closed due to their high risk profiles. The prerequisite is more difficult for the other list. The level 2 list is still sufficiently constrained for each individual situation to be considered. Credit and collections officers believed that at least half of the instances on this list may involve dubious fraudulent behaviour. The last and largest list is a challenging challenge. Only one-third of them seem suspect, though. For example, the query may contain the first five characters of passwords, email addresses, and phone numbers to improve time efficiency and save overhead expenses. The level 2 list and level 3 list could both be utilised with these additional questions.

V. RESULT

The code returns the quantity of false positives it identified after comparing the real numbers with that number. This is used to assess the algorithms' accuracy and precision.

The result is as follows, with class 0 designating a real transaction and class 1 indicating a transaction that was determined to be fraudulent. Along with the classification report for each method, these outcomes are also given.

This result was compared to the class values in order to rule out any potential false positives.

Logistic regression

Model summary

- Train set
  - Accuracy = 0.95
  - Sensitivity = 0.92
  - Specificity = 0.98
  - ROC = 0.99
- Test set
  - Accuracy = 0.97
  - Sensitivity = 0.90
  - Specificity = 0.99
  - ROC = 0.97

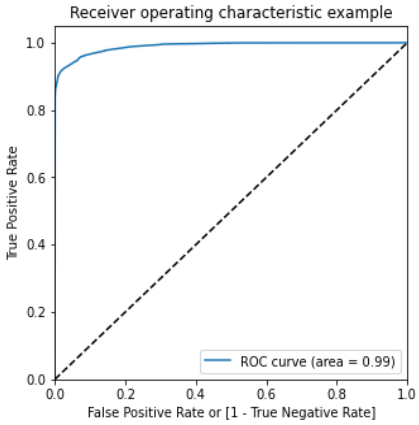

Fig. 1 ROC OF Train dataset using logistic regression

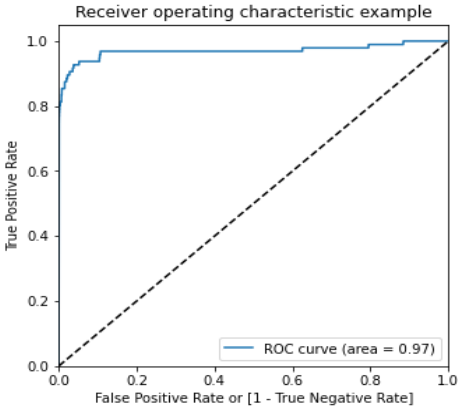

Fig. 2 ROC of Test dataset using logistic regression

Model summary

- Train set
- Accuracy = 0.99
- Sensitivity = 1.0
- Specificity = 0.99
- ROC-AUC = 1.0
- Test set
- Accuracy = 0.99
- Sensitivity = 0.79
- Specificity = 0.99
- ROC-AUC = 0.96

Overall, the model is performing well in the test set, what it had learnt from the train set.

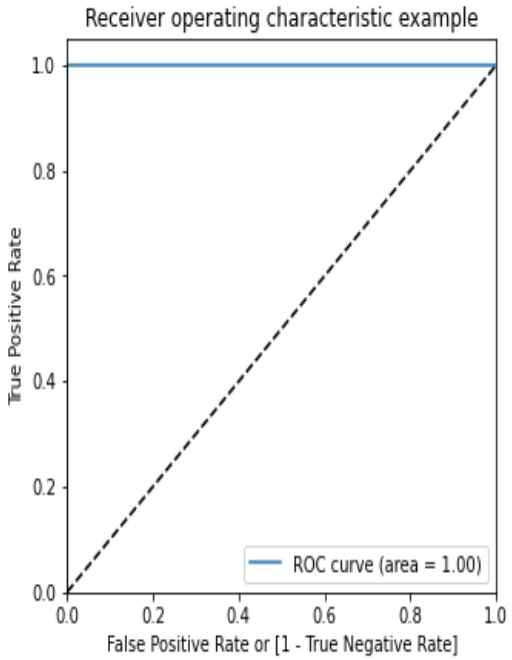

Fig.3 ROC of train dataset using xgboost algorithm

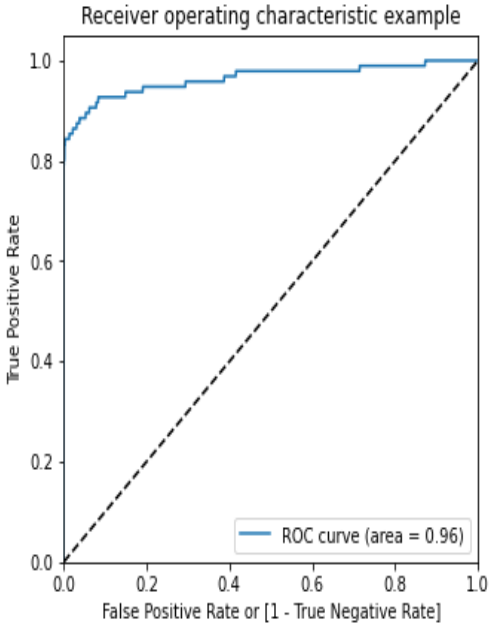

Fig. 4 ROC of Test dataset using xgboost algorithm

Decision Tree

Model summary

- Train set
  - Accuracy = 0.99
  - Sensitivity = 0.99
  - Specificity = 0.98
  - ROC-AUC = 0.99
- Test set
  - Accuracy = 0.98
  - Sensitivity = 0.80
  - Specificity = 0.98
  - ROC-AUC = 0.86

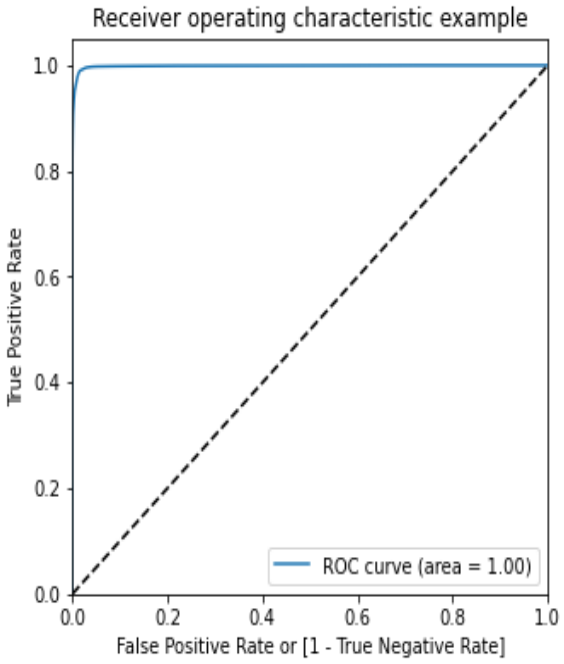

Fig. 5 ROC of train dataset using decision tree algorithm

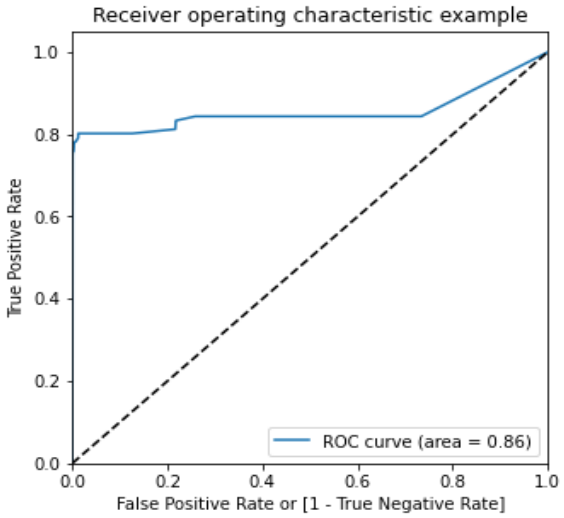

Fig. 6 ROC of test dataset using decision tree algorithm

Table 3 Classification report using decision tree algorithm

CONCLUSION

It goes without saying that using a credit card fraudulently is a crime. This page lists the most common fraud schemes and describes how to recognise them. It also discusses recent academic work in the field. This paper has also offered a full explanation of how machine learning may be used to enhance fraud detection along with the technique, pseudocode, description of how it is implemented, and outcomes of experiments.

Only a small piece of the dataset—which consists solely of just two days' worth of transaction records—could be made public if this study were to be applied on a commercial basis. The software will only get more efficient over time because it is built on machine learning principles.

The Logistic model can be considered as the best model to choose because of the easy interpretation of the models and also the resource requirements to build the model is lesser than the other heavy models such as Random Forest or XGBoost.

FUTURE ENHANCEMENTS

While we didn't achieve our goal of 100% accuracy in fraud detection, we did manage to create a system that, with more time and data, may get very close to it. As with any attempt of this kind, there is room for improvement here. The project's structure makes it feasible to integrate several algorithms as modules and combine their outputs to increase the accuracy of the final result. To improve this model even more, other algorithms may be included. However, the format of the output from these algorithms must match that of the others. Once that need is satisfied, as shown in the code, adding the modules is straightforward.

REFERENCES

[1] "Credit Card Fraud Detection Based on Transaction Behavior -by John Richard D. Kho, Larry A. Vea" was included in the proceedings of the 2017 IEEE Region 10 Conference (TENCON), which was held in Malaysia from November 5-8, 2017.

[2] CLIFTON PHUA, VINCENT LEE, KATE SMITH, & ROSS GAYLER are the authors. Published by the School of Business Systems, Faculty of Information Technology, Monash

University, Wellington Road, Clayton, Victoria 3800, Australia, "A Comprehensive Survey of Data Mining-based Fraud Detection Research"

[3] Research Scholar, GJUS&T Hisar HCE, Sonepat, "Survey Paper on Credit Card Fraud Detection by Suman," published in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 3, March 2014.

[4] Wen-Fang YU and Na Wang's "Research on Credit Card Fraud Detection Model Based on Distance Sum" was published by the 2009 International Joint Conference on Artificial Intelligence.

[5] By Massimiliano Zanin, Miguel Romance, ReginoCriado, and Santiago Moral, "Credit Card Fraud Detection using Parenclitic Network Analysis-By, Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages."

[6] AUGUST 2018 IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy"

[7] "Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, and Mridushi" appeared in the January 2016 issue of the International Journal of Advanced Research in Computer and Communication Engineering.

[8] "Plastic Card Fraud Detection Using Peer Group Analysis" Springer, Issue 2008. David J.Wetson, David J.Hand, M. Adams, Whitrow, and Piotr Jusczak.