Credit Card Fraud Detection Systems (CCFDS) using Machine Learning (Apache Spark)

Vinaya D S [1], Satish B Basapur[2], Vanishree Abhay[3], Neetha Natesh[4]

[1]M.Tech Student Information science & Dr. Ambedkar, Institute of Technology, Bangalore
[2,3,4]Asst.Professor Information science & Dr. Ambedkar Institute of Technology, Bangalore

---***---

Abstract *- As the payment method is simplified by the combination of the financial industry and IT technology, the payment method of consumers is changing from cash payment to electronic payment using credit card, mobile micropayment, and app card. As a result, the number of cases in which anomalous transactions are attempted by abusing e-banking has increased and financial companies started establishing a Fraud Detection System (FDS) to protect consumers from abnormal transactions. The abnormal transaction detection system aims to identify abnormal transactions with high accuracy by analysing user information and payment information in real time. Although FDS has shown good results in reducing fraud, but the majority of cases being flagged by this system are False Positives that resulting in substantial investigation costs and cardholder inconvenience. The possibilities of enhancing the current operation constitute the objective of this research. Based on variations and combinations of testing and training class distributions, experiments were performed to explore the influence of these parameters. In this study, we investigated the trend of abnormal transaction detection using payment log analysis and data mining, and summarized the data mining algorithm used for abnormal credit card transaction detection. We used python programming with apache spark for advanced processing of data and high accuracy.*

Key Words: *credit card, Fraud detection, Outlier detection, GBT classifiesr*

1. INTRODUCTION

Million and billions of people use the credit card for payment in both online and offline transaction, due to existence of widespread point of sale (POS).countless transaction occurred per minute everywhere in the planet. The reason behind fraud is negligence of user .when third person steal the most important information about credit card and user details easily fraud can be achieved. To detect what type of fraud occur during transaction, we need to face

Several challenges. Fetching that among all the transactions is occurred and which one is real could be a task.

Amongst the standard and very common ways of making payment globally and especially in North America, because of the presence of a far reaching point of sale. A huge number of individuals around the globe use charge cards to buy products and services by getting credit for a time of half a month. Any helpful framework could be mishandled and charge card is no exemption to this. Alongside the ascent of charge card use, extortion is on the ascent. Monetary Institutions (FIs) endure refined fake exercises and bear a large number of dollar misfortunes every year. In light of statistics [2] frauds account to more than $1 billion every year for Visa and MasterCard around the world.

Credit card companies and their part banks attempt to discover better approaches to forestall scams. A portion of the precautionary measures on the cards are magnetic stripes, 3D monograms, and CVC. Credit card companies are likewise taking steps to have alternate for credit cards as Smart Cards, be that as it may, in view of assessments this substitution will be over the top expensive because of the broad POS network in USA and the gigantic no. of cards available for use in those places. FIS additionally utilizing an assortment of computing mechanism, such as Neural Networks (NNs), to follow and distinguish dubious exchanges and ban them for additional examination.

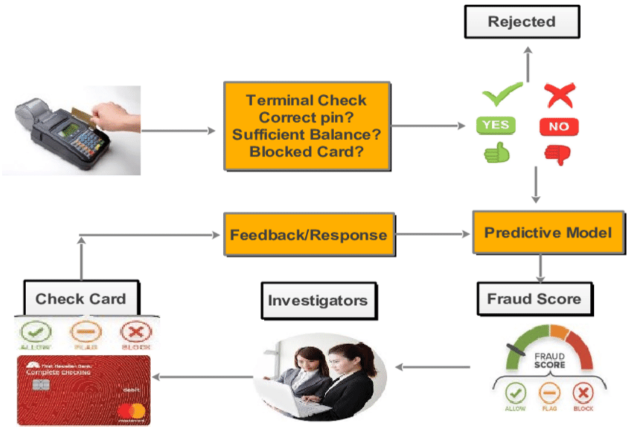


Figure 1: Card payment approval process

1.1Related work

Nagi et al. presented an intrusion detection frauds using data mining techniques for financial fraud detection. The papers from 49 journals published in 2018. This paper allows the analyzed and classified into four fraud categories and six data mining techniques [5].

Sanjeev et al. Is a paper that analyzes and classifies fraud type classification and fraudulent transaction frequency and amount by country through actual credit card fraud transaction data, and visually expresses the distribution using a box plot[6].

Michael et al. presented a signature based research on fraud detection and a fraud detection model [7] to provide a comprehensive survey of existing research related to fraud detection and to conduct fraudulent transactions in real time [7].

In the case of real-time detection, it is necessary to make accurate judgments in an instant and consider the characteristics of the data mining algorithm. TS Quah et al. Implemented real-time detection by separating detection mechanisms into the initial authentication layer, inspection layer, core layer for risk score evaluation and behavior analysis, and additional review layer using the SOM algorithm [17].

1.2 PROBLEM STATEMENT

Not al1 the doubtful transactions consider as fraudulent. It is commonly called as false positive (FP) which means that the case was not fraud although it was flagged as being potentially scam. This process of affirming each transaction those outliers from the cardholder's normal routine brings doubt about possible client disappointment. Additionally, the expenses related with exploring an enormous no. of false positives are high.

1.3 Motivation

As of now, a considerable amount of time is given for examining countless genuine cases (FPs). On the off chance that the quantity of examination on FPs could be dropped down, scam analysts can invest more energy and time in genuine fraud transactions that restricts the losses to the FIs.

2. SYSTEM ANALYSIS

2.1 Proposed System

The key objective of current research is improvising the procedure of personal follow up on a large number of suspicious transactions and to discover a path to preprocess the flagged records to recognize the probable genuine entries from the list of genuine/falsified entries. Here, the volume of needless analysis is decreased leading to significant savings for the financial institutions. Moreover, the current FDS threshold can also be lowered and a number of fraudulent cases, being missed under this level, can be detected. As a result, the fraud is discovered earlier and the overall losses may be reduced. For addressing these challenges, outlier detection and GBT Classifier is used, i.e. among the very common used applications of Machine Learning for addressing the pattern recognition and classification problems. The results indicate that the used method has a very good possibility to improvise the present system.

Advantages

- The proposed method overcomes the low accuracy forecast problem.
- Utilizing latest AI methods, the fraudulent transactions are recognized and the false alerts are reduced.
- Fast and reliable solution is attained.

2.1 Functional Requirements

Functional requirements are the characteristics of the product. All the features expected from any development are mentions as functional requirements.

2.2 Non-Functional Requirements

Non-Functional requirements list out the client expectations from product design, security, accessibility, and reliability or performance viewpoint.

Performance Requirements

Performance requirements tells about the software capability to respond on users' action such as:

- Upon running the application, it shouldn't take more than 3 seconds.
- Data validation shouldn't take above 5 seconds.
- Result generation should be achieved within 5 seconds.

Design Constraints-The project is to be developed in python which should get executed in a Windows OS. PyCharm editor should be used as an IDE.

Standards Compliance -There should be uniformity while defining variable names. The GUI shall have a poleasent look and feel. The graphical user interface should be user friendly.

Reliability-The product should not fail in mid of any operations carrying out.

Availability-The software can be used anytime.

Security-Security is very important for any application that holds user sensitive data.

Maintainability-The software admin should be able to manage the data.

Portability-The project should be executable on any Windows OS.

2.3 System Requirements:

Hardware Requirements:

Processor : Pentium i3 or higher.

RAM : 4 GB or higher.

Hard Disk Drive : 20 GB (free).

Peripheral Devices : Monitor, Mouse and Keyboard.

Software Requirements:

Operating system : Windows 8/10.

IDE Tool : PyCharm

Coding Language : Python 3.6

APIs : Numpy, Pandas,PySpark, Matplotlib

3. SYSTEM DESIGN

3.1 System Architecture:

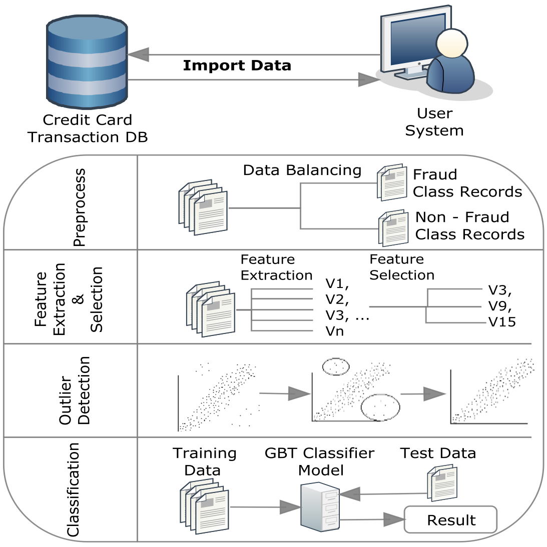


Figure 2: System Architecture

Above fig shows the process of CCFDS. This system model accepts real time customer credit card transaction database.it is more important to find fraud rate of credit card.

Data collection: collect input dataset based on transaction details,

Data balancing: after collecting large set of database it is necessary to understand and separate the balanced data and unbalanced data in two types of class .clas-0 indicates non-fraud and class-1 indicates fraud.

Feature extraction and selection: class-1 indicates total fraud transactions are 492 samples. In this project v1, v2 …v28 features.

Outlier detection: It measures the distance between each similar data to the clustering technique. The values which are not follows the trained data consider as outlier.

Classification: As the dataset is imbalanced, many classifiers show bias for majority classes. PySpark library is applied as a SQL-like analysis to a large amount of structured or semi-structured data. GBT Classifier does the classification of data coming through the stream.

3.2 Use Case diagram

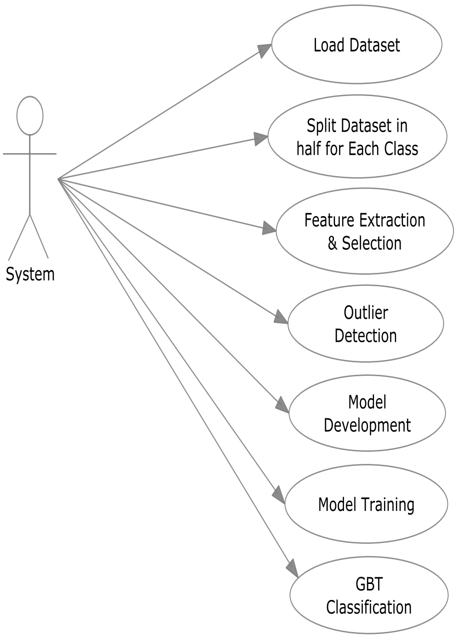


Figure 3: Use Case Diagram

3.3 DFD (Data Flow Diagram)

The DFD used as communication tool between system and user.it is a simple representation of the complete project process. Transaction detection activity follows three phases.1.Data exploration 2.Data prepressing 3.data classifications.

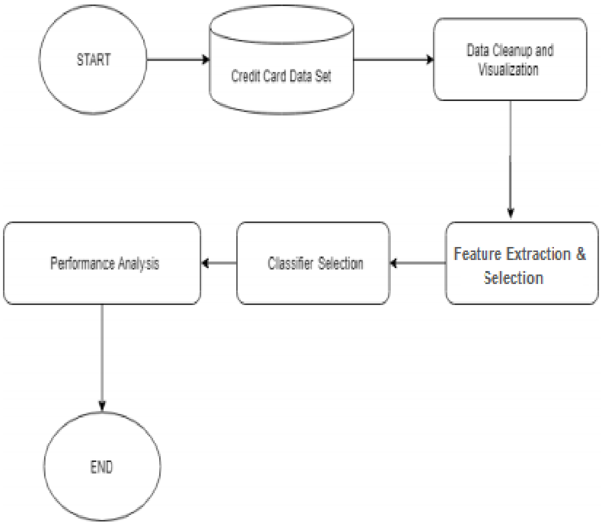


Figure 4: DFD Level 1

4. SYSTEM IMPLEMENTATION

The proposed system is divided into several smaller units, known as module. The main functional modules of the proposer system are given below.

4.1 Module Names

1. Data Collection
2. Data Balancing
3. Feature Extraction
4. Outlier Detection
5. Classification

4.2 Module Description

1. **Data Collection:** It contains 2,84,807 records of credit card transactions that happened in duration of just 2 days. This dataset is very much unbalanced as it has a total of 492 fraud entries 2,84,515 genuine entries i.e. is just 0.17% of total records. The original features are masked with V1, V2, V3, ...V28. The last column here represents fraud or non-fraud class i.e. represented by 0 and 1 respectively.

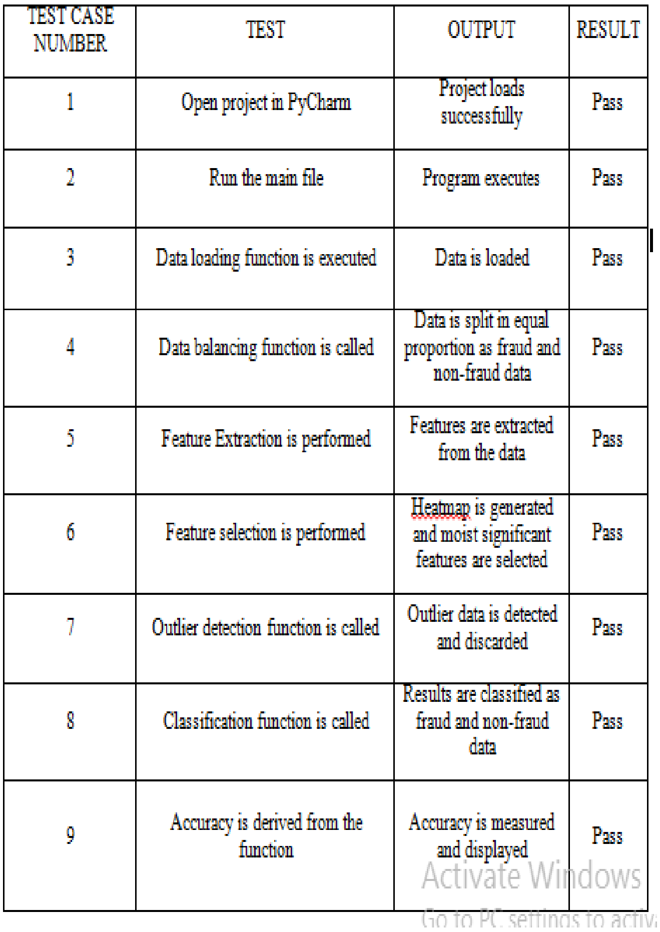
2. **Data Balancing:** Imbalanced classes are a general issue in ML based classification where there is an abnormal count of each class. It occurs due to the fact that ML Algorithms are typically intended to improve precision by diminishing the errors. In this manner, they don't consider the class or balancing the ratio of classes. As out of 2.84807 transactions just 492 fraud transactions exist, which makes it quite difficult to build a standard model with this much less number of fraud transactions. Thus, we use pandas in python to make it 50-50 i.e. we decrease the no. of legitimate transactions to balance it with the number of fraud transactions in equal proportion.

3. **Feature Extraction:** We use heatmap technique to find the significant feature that can distinguish the classes properly and ultimately that affects the accuracy of detection algorithm. Heatmap provides a good visualization of the major and minor values in the matrix as different colored cells that define the values. Here, rows/columns of the matrix are clustered in sets. Thus, the features which look most significant are recognized and used further for model training.

4. **Outlier Detection**: The outlier detection technique measures the distance of each data similar to the clustering technique, but is used to find specific data and rules that are separated from the total data. The values which are not in flow of the linear graph are considered as outliers. Here our aim is to reduce the outliers to have a better trained model. We use numpy library in python for this.

5. **Classification:** The task of classification occurs in a wide range of applications. In a broad sense, the term could relate to any context in which some decision or forecast is made on the basis of currently available information. It works on a set of pre-defined classes on the basis of observed attributes or features. Here the aim is to establish a rule whereby one can classify a new observation into one of the existing classes. The construction of a classification procedure from a set of data for which the true classes are known has also been variously termed as pattern recognition, discrimination, or supervised learning. We use PySpark and GBT Classifier for data streaming and classification purpose. PySpark library is applied as a SQL-like analysis to a large amount of structured or semi-structured data. GBT Classifier does the classification of data coming through the stream.

5. TEST CASES



TEST CASE NUMBER	TEST	OUTPUT	RESULT
1	Open project in PyCharm	Project loads successfully	Pass
2	Run the main file	Program executes	Pass
3	Data loading function is executed	Data is loaded	Pass
4	Data balancing function is called	Data is split in equal proportion as fraud and non-fraud data	Pass
5	Feature Extraction is performed	Features are extracted from the data	Pass
6	Feature selection is performed	Heatmap is generated and moist significant features are selected	Pass
7	Outlier detection function is called	Outlier data is detected and discarded	Pass
8	Classification function is called	Results are classified as fraud and non-fraud data	Pass
9	Accuracy is derived from the function	Accuracy is measured and displayed	Pass

6. RESULT AND ANALYSIS

Step 1: Install the required software and files like PTHON 3.6 and pycham IDE

Step 2: Create a folder to load credit card database

Step 3: Open pycharm ide

Step 4: Create a new project

Step 5: Import the credit card database to new project

Step 6: Write executable credit card fraud detection program code

Step 7: Verify and debug the code to get better and good result

Step 8: Run the project code

Step 9: Check he final result

------------------*******-----------------------------

Total records : 284807

Fraud record : 492

Non-fraud record: 284315

True positive : 79

True negative : 77

False positive : 4

False negative : 6

Recall : 0.9294117647058824

Precision : 0.9518072289158826

Accuracy : 92.9411764

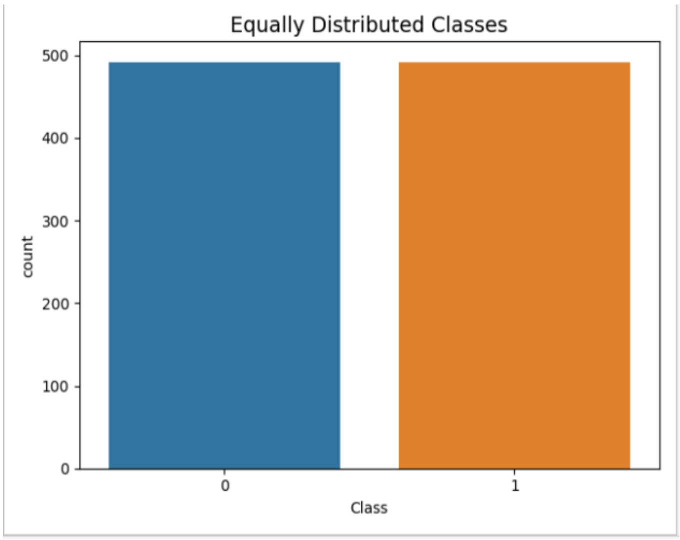


Fig5: Distributed class

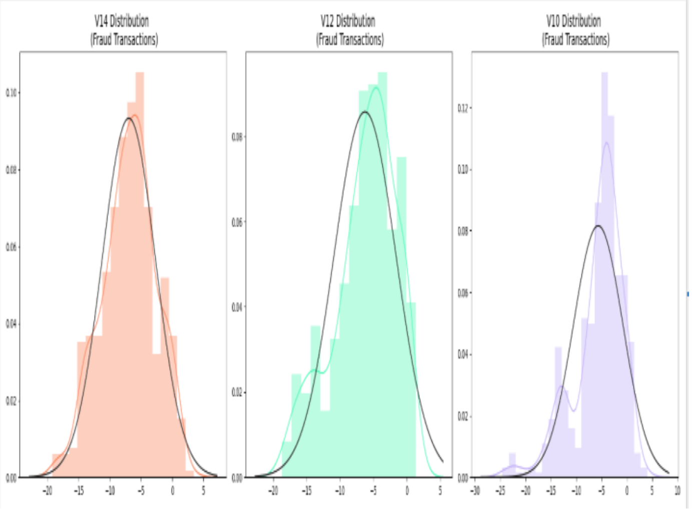


Fig6: Fraud transaction detection of features

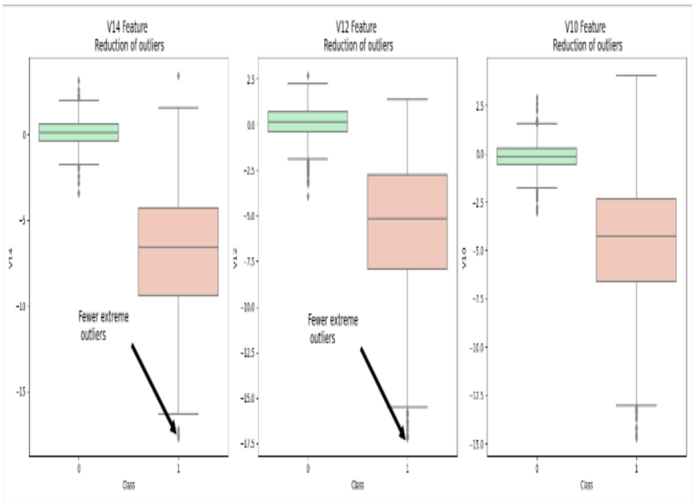


fig 7:Feature reduction of outliers

7. CONCLUSIONS

With the development of electronic financial transaction technology and the emergence of simple payment, the risk of fraudulent payment and fraudulent payment increases as the authentication process is simplified. The types of fraudulent use of credit cards include theft and loss, identity theft, new card not received, card forgery, and card information theft. In particular, as phishing, pharming as well as card information leakage due to card information leakage, card information theft accidents are occurring. In response, the government tried to deal with electronic financial fraud by implementing the 'e-financial fraud prevention service'. It is difficult to cope with financial fraud by simply setting the existing keyboard security, public certificate, and additional password. The abnormal transaction detection system is used to analyze the user's data and payment data in real time to inform the financial institution and the user of the detection if it is different from the usual pattern, and further to arbitrarily stop the transaction. Therefore, an abnormal transaction detection system is important for fast and accurate detection, and research is needed to improve the algorithm. In this study, the method of detecting anomalous transactions using the electronic payment log analysis and machine learning technique was investigated. Results show the significance of algorithms used over the dataset and efficient classification is performed.

In future deep learning concepts can be applied using convolution networks for improved accuracy. Also some other datasets can be used for further testing of proposed mechanisms.

REFERENCES

[1] Donald V. Macdougall, Richard G. Mosley, Garioch J. l. Saunders; Credit card crime in Canada: Investigation - Prosecution; The Canadian Association of Crown Counsel; page 1-56; January 1985.

[2] Isabelle Sender; Detecting and combating fraud; Chain Store Age; New York; Vol. 74; Issue 7; Page 162; July 1998.

[3] Elford Dean, Raj Thomas, Lorry; Visa security center; Personal meetings; January 7 and February 11,1999.

[4] Gyusoo Kim and Seulgi Lee, "2014 Payment Research", Bank of Korea, Vol. 2015, No. 1, Jan. 2015.

[5] EWT Nagi, Yong Hu, HY Wong, Yijun Chen, Xin Sun, "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature," Decision Support Systems, Vol. 50, No. 3, Feb. 2011.

[6] Jha, Sanjeev, J. Christopher Westland, "A Descriptive Study of Credit Card Fraud Pattern," Global Business Review, Vol. 14, No. 3, pp. 373-384, 2015.

[7] Edge, Michael Edward, Pedro R. Falcone Sampaio, "A Survey of Signature based Methods for Financial Fraud Detection," Computers & Security, Vol. 28 No. 6, pp. 381-394. 2009.

[8] Aihua Shen, Rencheng Tong, Yaochen Deng, "Application of Classification Models on Credit Card Fraud Detection," Service Systems and Service Management of the 2007 IEEE International Conference, pp. 1-4, Jun.2007.

[9] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi, Hao Fu, "Financial Fraud Detection Model: Based on Random Forest," International Journal of Economics and Finance, Vol. 7, No. 7, pp. 178-188, 2015.

[10] Ganesh Kumar.Nune and P.Vasanth Sena, "Novel Artificial Neural Networks and Logistic Approach for Detecting Credit Card Deceit," International Journal of Computer Science and Network Security, Vol. 15, No. 9, Sep. 2015..