Fraud Detection using Isolation Forest Algorithm in Credit Card

Author Name- Mrs. Vandana Tripathi(vandy3585shi@gmail.com)

Divyanshu Mishra(divyanshumishra45186@gmail.com)

Dhruv Gupta(dg739390@gmail.com)

Prakhar Tiwari(t.prakharofficial@gmail.com)

Department of Computer Science, Babu Banarasi Das Institute of Technology and Management, Lucknow, India

Abstract: Nowadays, everyone uses credit cards for a variety of transactions because of the demonetization of the economy. Therefore, there will be a higher likelihood of fraud. Banks maintain several, vast datasets. These data repositories can be used to extract crucial business data. Fraud is a problem having wide-ranging effects on the financial sector, the government, corporate sectors, and regular consumers. In recent years, the issue has arisen with an increase in reliance on new technologies like cloud and mobile computing. Physical detections are not only expensive and time-consuming, but they also don't produce reliable results. It is hardly unexpected that financial and computationally based automated processes have been adopted by economic institutions. Traditional methods relied on labor-intensive manual methods like auditing, which arebecause of the complexity of the issue, inefficient and unreliable. The capacity of data mining-based systems to spot minute anomalies in huge data sets has demonstrated its value. So, in order to detect fraud, we have employed several supervised algorithms, which produce reliable findings. There are numerous sorts of fraud and various data mining techniques, and study is always being done to determine the optimal strategy for each situation. Although the phrase "financial fraud" has several potential connotations, for our purposes, it can be defined as the intentional use of illegal means or means of doing things in order to earn financial advantage.Fraud has a significant detrimental effect on society and industry; just credit card fraud alone results in billions of dollars in annual revenue loss. For the classification of authentic and fraudulent records from the dataset, we used isolation forests.

Keywords: Fraud detection, Decision tree, Isolation Forest Algorithm

I.INRODUCTION

Fraud is the misappropriation of a profit-making organization's system without necessarily resulting in overt legal issues. The general act of deceiving another person or organisation in order to gain financial gain is fraud. Detecting credit card fraud involves classifying fraudulent transactions into two categories: legitimate transactions and hoaxes. The classic card-related frauds, internet frauds, and other types of fraud can all be generally grouped into three groups. Customer fraud and external fraud are terms used to describe fraud performed by people outside of the business, whereas management fraud and internal fraud are terms used to describe fraud committed by top-level management.Fraud detection automates and helps decrease the manual components of a screening

process because it is a part of comprehensive fraud control.

A credit card is a little, handy piece of plastic that carries identity data, such as a signature or photo, and allows the holder to charge goods or services to his account, for which he will receive recurring bills. Today, automated teller machines (ATMs), store readers, banks, and online internet banking systems all read the information from credit and debit cards. They have a special card number, which is crucial. Both the physical security of the plastic card and the confidentiality of the credit card number are essential to its security. The quantity of credit card transactions is increasing quickly, which has caused a significant increase in fraudulent activity.Using a credit card as a dishonest source of funds in a transaction is referred to as credit card fraud, which covers a wide range of theft and fraud offences. To handle this fraud detection problem, statistical approaches and several data mining algorithms are typically applied. Artificial intelligence, meta learning, and pattern matching are the main foundations of most credit card fraud detection systems. The goal of genetic algorithms, which are evolutionary algorithms, is to find the best solutions for removing fraud. The development of effective and secure electronic payment systems is given top priority in order to determine if a transaction is fraudulent or not. This essay will concentrate on credit card fraud and the methods used to catch it.

Types of algorithms and graphical representations of them. [ZivileGrundiene, Rasa Kanapickiene] "The financial ratios-based fraud detection methodology" This author describes how financial ratios are analysed to identify the financial ratios of financial statements that are most susceptible to fraud in light of the motive of business managers and employees to commit fraud. It was discovered that fraud is typically performed to demonstrate the company's continued growth and to satisfy contractual obligations. Such ratios can be found in a wide variety in literary sources. According to theoretical research, profitability, liquidity, activity, and structural ratios are the most frequently examined ratios.A theoretical investigation found that financial ratios are examined in scientific literature to determine which ratios of the financial statements are the most susceptible to executive managers' and employees' motivations for committing fraud. The financial

II.LITERATURE SURVEY

[Maumita Bhattacharya, Jarrod West] Detecting intelligent financial fraud Although the performance of each strategy varied, they were all shown to be reasonably capable of spotting various types of financial fraud. This author discusses numerous clever statistical and computational techniques to fraud detection. The capacity of computational approaches like neural networks and support vector machines to pick up and adjust to a wide variety of new strategies is extremely useful to the tactic evolution of fraudsters. Initial studies on fraud detection largely on statistical models like logistic regression and neural networks. Forecasting is one financial application where neural networks are applied. Neural networks have a long history of use in fraud detection.However, they are unsuited for real-time function since they require a lot of processing power for both training and operation. If the training set is not a suitable representation of the issue domain, there is a risk of over fitting, necessitating continuous retraining to accommodate new fraud techniques. The author discusses various types of fraud in this essay, including insurance fraud, mortgage fraud, health insurance fraud, telecommunications fraud, and credit card fraud. For various fraud types, many strategies have been developed, with factors like entropy and sensitivity established and the effectiveness of the various techniques compared. statements fraud detection logistic regression model has been built. [Surya B. Yadav, Fletch H. Glancy] A computer model for fraud detection in financial reporting The content of yearly filings with the Security and Exchange Commission can be used by the computational fraud detection model to identify financial exposure fraud, according to this author. The concept is adaptable to different disciplines and genres because it outlines automatable procedures.To screen businesses for prospective SEC fraud investigations is one possible use for CFDM (Security and exchange commission). Financial analysis, email spam identification, and business intelligence validation are further potential applications. For the purpose of identifying fraud in financial reporting, a computational fraud detection model (CFDM) was put forth. On textual data, CFDM employs a quantitative method. It uses

methods for fraud detection that effectively make use of all the information in the textual data.

III.PROPOSED METHODOLOGY

There are several challenges that make this process difficult to accomplish, but one of the main issues with fraud detection is the dearth of real-world data for academic researchers to conduct experiments on, as well as experimental literature that provides real-world outcomes. The sensitive financial information related to the fraud that must be kept secret in order to protect the privacy of the consumer is the cause of this. Here, we list some qualities that a fraud detection system should possess in order to produce accurate results: (Since only a very small portion of all credit card transactions are fraudulent, the system ought to be able to manage skewed distributions. There must to be a suitable way to deal with the noise.Data mistakes, such as inaccurate dates, are referred to as noise. Regardless of how big the training set is, the noise in the real data restricts how accurate generalisation can be.

(Overlapping data is another issue in this subject. Many transactions could appear to be fraudulent when they're actually legitimate. When a fraudulent transaction looks to be legitimate, the inverse also occurs.

The systems ought to be able to adjust to new varieties of fraud. Since an effective fraudster is continuously looking for new and creative ways to do his work, successful fraud techniques eventually lose some of their effectiveness as they become publicly recognised.

Good metrics are required to assess the classifier system. For instance, even with a very high accuracy, practically all fraudulent transactions can be misclassified, hence the total accuracy is not suitable for evaluation on a skewed distribution. The system should account for both the amount of money lost to fraud and the amount that will be needed to uncover that fraud. For instance, it is not profitable to stop a fraudulent transaction that will cost far less to identify than it will cost to stop it.

Isolation Forest

By selecting a feature at random, followed by a split value between the maximum and minimum values of that feature, the Isolation Forest "isolates" observations.

The number of splittings necessary to isolate a sample is equal to the length of the path from the root node to the

terminating node since recursive partitioning can be represented as a tree structure.

This path length provides a gauge of our decision function and normalcy when averaged over a forest of similar random trees. For anomalies, random partitioning results in considerably shorter pathways. As a result, shorter path lengths for specific samples produced by a forest of random trees are quite likely to be outliers.

Figure 1 from [1] illustrates the concept of distinguishing a normal observation from an aberrant observation. In comparison to an anomalous point, a normal point (on the left) requires more partitions to be found (right).

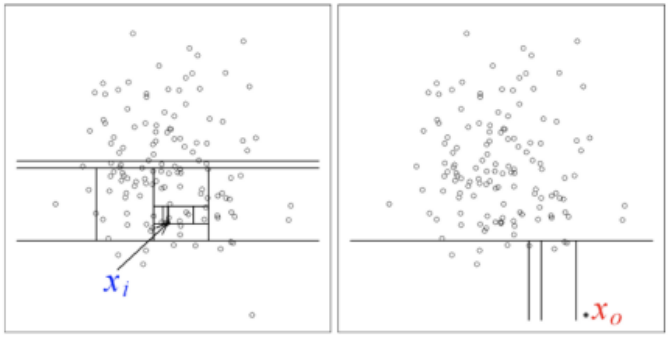

Figure 1 Identifying normal vs. abnormal observations

For decision-making, an anomaly score is necessary, much like with other outlier detection techniques. It is defined as follows for Isolation Forest:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where c(n) is the average path length of unsuccessful searches in a binary search tree, n is the number of external nodes, and h(x) is the path length of the observation x. You may learn more about the anomalous score and its elements in [1]. Every observation is given an anomaly score, and based on that score, the following choice can be made:

(Scores near 1 denote oddities.

(A score substantially lower than 0.5 denotes typical observations.)

If all scores are near to 0.5, then there do not appear to be any obvious anomalies over the entire sample.

IV. CONCLUSION

In recent years, credit card fraud has increased dramatically. Building a precise and user-friendly credit card risk monitoring system is one of the main challenges for the merchant banks in order to raise the level of risk management for merchants in an autonomous and efficient manner. Finding the user model that best detects fraud incidents is one goal of this study. Credit card fraud can be found in a variety of ways. Bank credit card fraud detection systems can estimate the likelihood of fraudulent transactions shortly after credit card transactions if one of these algorithms, or a mix of them, is used. Additionally, a number of anti-fraud methods can be used to lower risks and stop institutions from suffering significant losses.This research contributes to the supervised learning algorithms used in credit card fraud detection. Using Isolation Forest, we were able to find outliers with a 99.67% accuracy rate for all classified records.

IV.MODELING & ANALYSIS

IV.I REQUIREMENTS

The first stage, commonly known as the data exploration stage, entails loading the dataset. We have utilised visual exploration to learn what is in the dataset and its characteristics. Data exploration is a procedure similar to data analysis. We used a data set from the Kaggle website that was reduced in size using the PCA dimensionality reduction procedure and comprises numerous characteristics including quantity, class, time,

Software specifications:(Used Program: Python 3.6.5 (Jupyter Notebook) (Operating System: Windows 10) (x86)

Data set Requirements:

(Credit card dataset) requirements (.csv file containing 2,84,807 records)

IV.II. PROJECT PHASES

We are conducting the following three steps of this fraudulent transaction detection activity:

1)Steps for Data Exploration

a) load the dataset; b) preprocess the dataset;

c) make a graph; d) display the dataset.

2) Steps in Data Preprocessing

A) Load DatasetB) Remove Null ValuesC) Split Dataset D) Advance to Training Phase

3) Steps for Data Classification

A) Train the dataset B) Create a classifier C) Create an Isolation Forest D) Perform Classification

and others. To provide descriptive statistics that capture the central tendency, dispersion, and form of a dataset's distribution for the specified series object, the dataset is investigated and represented. Every calculation is done without include Null values. The histogram is created by exploiting this information to represent it.

The preprocessing of data is the second stage. To increase its effectiveness, the dataset is reloaded and all null values and garbage values are eliminated. We must divide the dataset into training and testing phases within this phase itself. By describing class 0 as a valid transaction and class 1 as a fraudulent one, we are primarily working on the training phase here. being

trained To improve the quality and produce more accurate data, the dataset is offered at random along with both legitimate and fraudulent entries. In order to summarise the data, provide an input for a more sophisticated study, and serve as a diagnostic for that analysis, a correlation matrix is offered.

The data classification phase is the third and last one. It just involves entering training data sets with pre-labeled classes for the algorithm to learn from. The model is then applied by supplying a separate dataset for which the classes are not defined, and using the knowledge gained from the training set, the model guesses the class to which it corresponds. To get a useful result identifying The two algorithms must be applied. employing terms like accuracy, recall, f1-score, and support as an outcome.

V.RESULTS & DISSCUSSIUONS

V.I.CORRELATION MATRIX

A heat map called the correlation matrix is used to determine whether there is any link between various parameters and various variables in our dataset.

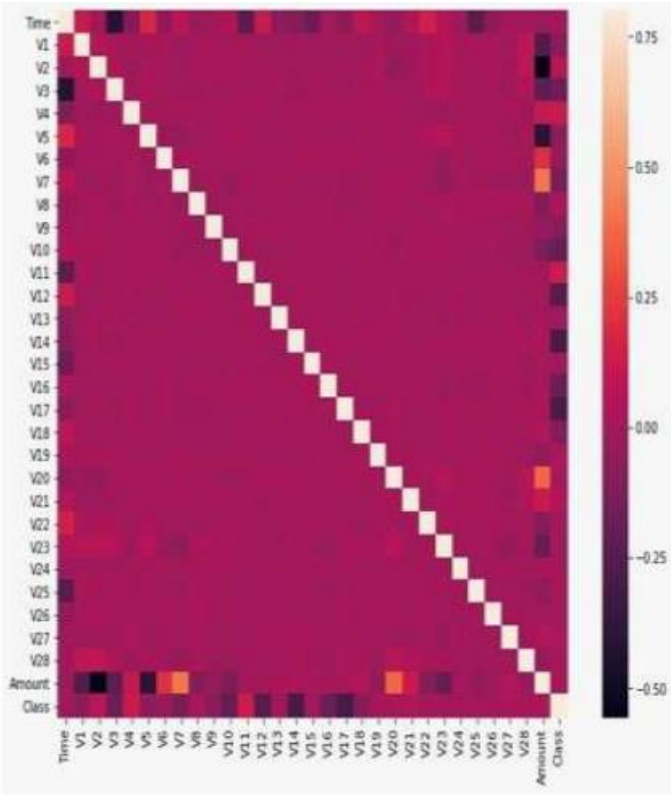

Fig -4.1.: Heat Map

The heatmap for SNS and seaborn are used in the pyplot figure above. It offers our straightforward correlation matrix a visual appearance and facilitates analysis. Both the X and Y axes contain all 31 parameters (V1 through V29), class, and amount, with a range of -0.75 to +0.50.

V.II.HISTOGRAMS

The project makes use of histograms to quickly analyse genuine and fraudulent transactions. The matplotlib can be used for this. Additionally, we can adjust the plot's size accordingly.

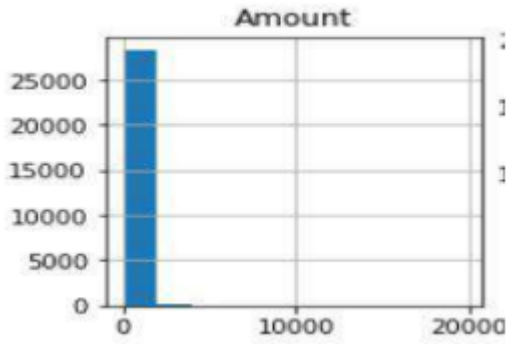

Fig -4.2.: Bar graph for amount

The aforementioned histogram displays the average amount spent by all clients as well as the proportion of legitimate and fraudulent transactions. A transaction is considered to be real if it is at 0, and fraudulent if it is at 1. According to the graph, 98 percent of the transactions were legitimate, while only 2 percent were deemed fraudulent. In this method, we can use straightforward histograms to assess all 31 parameters.

V.III. PERFORMANCE METRICS

The effectiveness of various machine learning algorithms can be assessed using a wide range of indicators. We have just used the major four measures, though.

Tp: true positive values

Tn: true negative values

Fp: false-positive values

Fn: false negative values

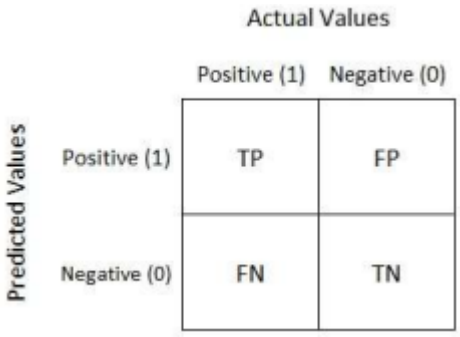

Fig -4.3.: Confusion metrics

These performance indicators are computed using the confusion matrix. The following terms are those used in the confusion matrix:

• Accuracy: It is the quantitative relationship between Tp and (Tp + Fp). Zero is the worst possible value while one is the finest.

• Recall: It is Tp/(Tp+Fnquantitative)'s relationship. It locates every potential positive sample.

• F1 score: This score represents a weighted average of recall and precision.

Support: It is the proportion of true values to each goal value's number.

VI. CONCLUSIONS

One of the main problems with contemporary online transactions is the detection of credit card fraud. It is essential to create a classifier that can distinguish between fraudulent and legitimate transactions in order to prevent such a situation from occurring. To integrate further processes, we will split the dataset into training and testing phases. For the training phase, we need a dataset that contains both fraudulent and real entries. In order to detect fraudulent transactions and maintain the legitimacy of the payment system, machine learning algorithms (Isolation forest and local outlier factor) that can best adapt to the change in scenario taking place can be utilised and created on a very large scale.

VII. REFERENCES

[1] Vladimir Zaslavsky and Anna Strizhak," credit card fraud detection using self organizing maps", information & security. An International Journal, Vol.18,2006.

[2] Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), "Credit card fraud and detection techniques: a review", Banks and Bank Systems, Volume 4,Issue2,2009.

[3] Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, "A review of Fraud Detection Techniques: Credit Card", International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012 .

[4] L. Mukhanov, "Using bayesian belief networks for credit card fraud detection," in Proc. of the IASTED International conference on Artificial Intelligence and Applications, Insbruck, Austria, Feb. 2008, pp. 221– 225.

[5] Linda Delamaire ,Hussein Abdou and John Pointon, "Credit Card Fraud and Detection technique", Bank and Bank System,Volume 4, 2009.

[6] John T.S Quah,MSriganesh "Real time Credit Card Fraud Detection using Computational Intelligence" ELSEVIER Science Direct,35 (2008) 1721-1732.

[7] Joseph King –Fung Pun, "Improving Credit Card Fraud Detection using a Meta Heuristic Learning Strategy" Chemical Engineering and Applied Chemistry University of Tornto 2011.

[8] Kenneth Revett,Magalhaes and Hanrique Santos "Data Mining a Keystroke dynamic Based Biometric Dtatabase Using Rough Set" IEEE.