INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact Factor: 6.078

(Volume 7, Issue 4 - V7I4-1435)

Available online at: https://www.ijariit.com

Credit Card Fraud Prediction

Thyagaraj Tanjavur
thyagaraj_tanjavur@bmsit.in
BMS Institute of Technology and Management, Bengaluru, Karnataka

Yash Rajesh
1by17ec183@bmsit.in
BMS Institute of Technology and Management, Bengaluru, Karnataka

ABSTRACT

Fraud detection by credit companies is essential in this digital era where majority of the financial transactions are made online. Fraudsters use loopholes in the payment systems to their benefit. Such problems can be solved to a large extent if the companies add an extra layer of security before confirming the transactions using machine learning algorithms. This project intends to use Isolation Forest algorithm to enhance the security of credit card transactions by predicting the credibility of the transaction before authorization. Detecting 100% of the fraudulent transaction, minimizing the incorrect fraud classifications and making the process automated is our objective.

Keywords: *Credit card fraud, Extra security layer, Machine learning systems, Isolation Forest algorithm Automated fraud prediction*

1. INTRODUCTION

Credit card fraud includes frauds committed using a payment card, such as a credit card or debit card. The purpose can be obtaining goods or services, making payment to another account, etc. The Payment Card Industry Data Security Standard (PCI DSS) is the data security standard that helps businesses process card payments securely and reduce frauds.

Credit card fraud can be authorised, where the genuine customer themselves processes a payment to another account which is controlled by a criminal, or unauthorised, where the account holder does not provide authorisation for the payment to proceed and the transaction is carried out by a third party.

Credit card frauds are a very alarming issue in India. After a rise in credit card frauds, the RBI had asked banks to add security features including an electronic chip and a secret PIN which the cardholder is required to punch in the terminal to authenticate payment. This paper proposes a method using machine learning to predict frauds even before transavtion is authorized by banks.

Fraud prediction involves analayzing the previous transaction history of the user and using it to predict the nature/behaviour of the next transaction initiator whether it matches with the real user or not.

In real world application of this system, co-operation from the banks and behavioral changes if the user may pose some problem and an extra verification step may be tiring for the users. But nevertheless the risks vs reward factor makes the proposed system very beneficial in avoiding frauds.

1.1 Flowchart

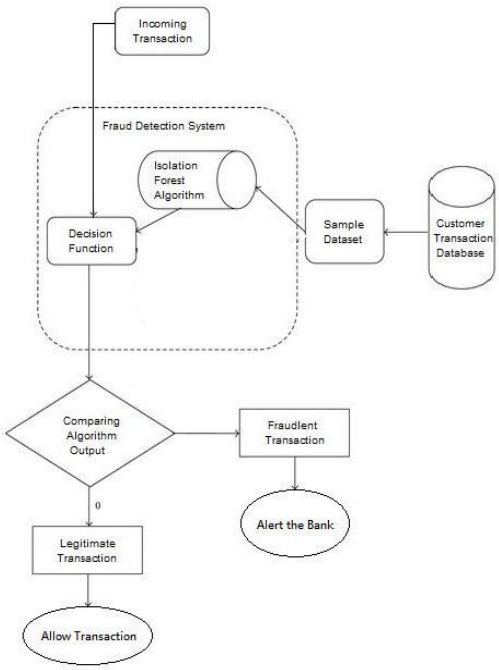

Chart-1 Flowchart of proposed system

2. LITERATURE REVIEW

Fraud is an unlawful or criminal deception for financial or personal benefit. It is voluntary act against the law, rule or policy with to attain unauthorized financial benefits. Many different papers pertaining to anomaly or fraud detection in this domain have been published already and are available for

public usage. A difference in existing and proposed system is given the following image.

Table 2.1 Existing system vs. Proposed system

EXISTING SYSTEM	PROPOSED SYSTEM
1. In existing system methods such as Cluster Analysis, SVM, Bayesian network, Logistic Regression, Naïve Bayer's, Hidden Markov model etc are used to find out the credit card fraud transactions.	1. In proposed System, we use Random forest algorithm to classify the credit card dataset. Random Forest is an algorithm for classification and regression.
2. The methods used in the existing system are based on unsupervised learning and the accuracy obtained by these methods is about 60-70%.	2. Even for large dataset this algorithm is extremely fast and can able to give accuracy of about 98%. Finally the number of fraud transactions will be identified and represented in the form of confusion matrix.

3. METHODOLOGY

This paper proposes the detection of anomalies/outliers using isolation forest algorithm.

The implementation can be viewed as:

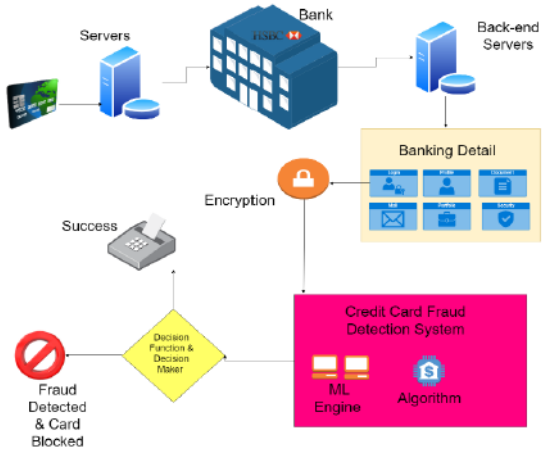

Fig-1 Real time implementation architecture

Our dataset was obtained from Kaggle, a data analysis website for data scientists that provides free datasets to work on. This dataset contains 31 columns and 28 of it (indicated as v1-v28) are protected sensitive information. The remaining three columns represent Amount, Class and Time. Time indicates the time interval between the first transaction and the successive transaction. Amount indicates the amount of money transacted. Class 0 defines a valid transaction and 1 defines a fraudulent one. Then, to visually comprehend the data we'll plot different graphs:

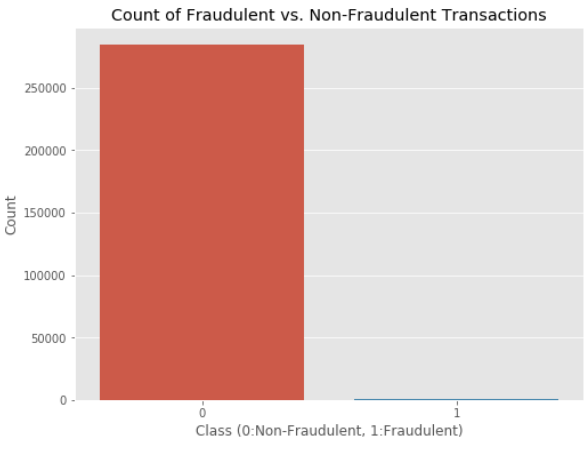

Fig-2 Count of Fraudulent vs Non-Fraudulent Transactions

This shows that the number of non-fraudulent transactions is much higher compared to fraudulent ones.

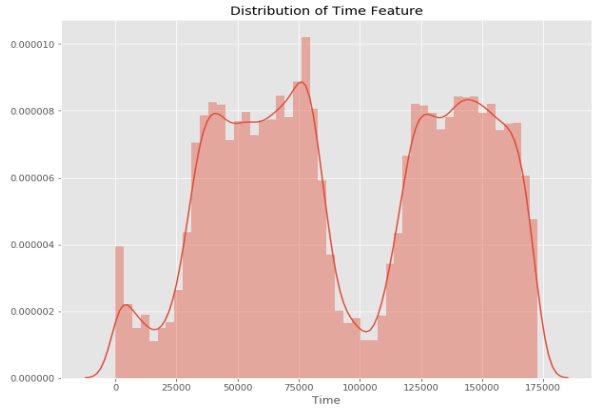

Fig-3 Distribution of Time of transactions

This shows that majority of the transactions were made during day time.

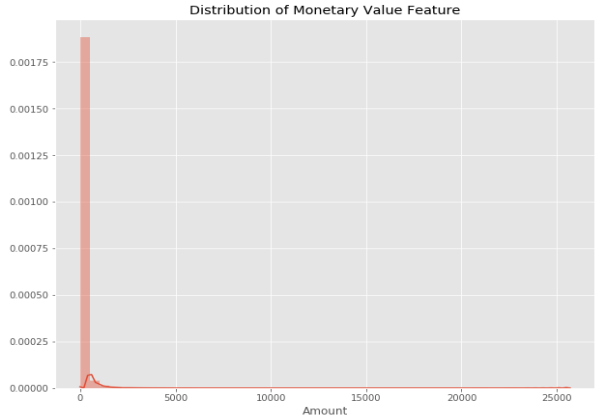

Fig-4 Distribution of Amount transacted

This shows that majority of the transactions had low amounts. We then plot and use a histogram for every column to ensure that there are no values missing in the dataset because we don't require any missing value imputation and the machine learning algorithms can process the dataset smoothly.

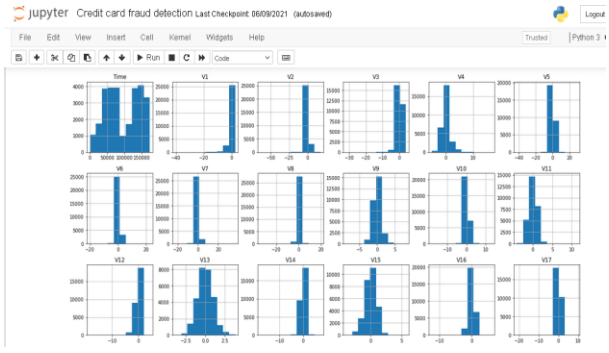

Fig-5 Histogram for columns v18-28, amount and class

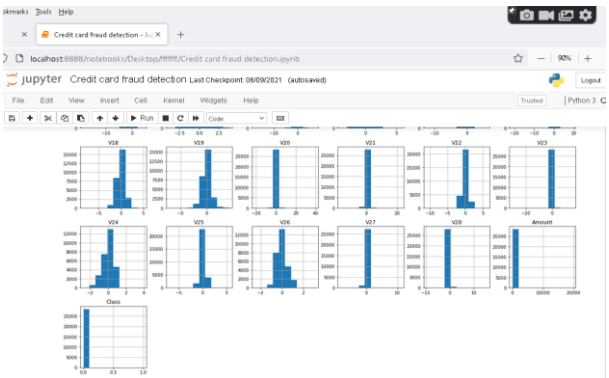

Fig-6 Histogram for columns v18-28, amount and class

After this analysis, to study the correlation between our predicting variables and the class variable, we plot a heatmap.

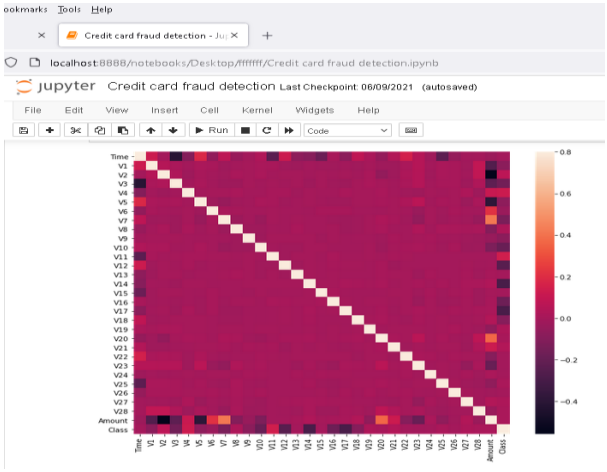

Fig-7 Heatmap

After formatting and processing of data, we now use the Isolation Algorithm on the Amount column to detect the outliers and assign a score to them. Based on the score, anomaly score will be assigned to each row (1 for genuine transactions and -1 for anomalies).

3.1 Isolation Forest Algorithm

The Isolation Forest isolates the outliers. It selects a feature randomly and then selecting a split value between the maximum and minimum values of the selected feature randomly. Recursive partitioning can be represented by a tree structure and the number of splitting required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality and our decision function. Such random partitioning produces shorter paths for anomalies. Hence, when a forest of random trees together produce shorter path lengths for particular samples, they are highly likely to be outliers.

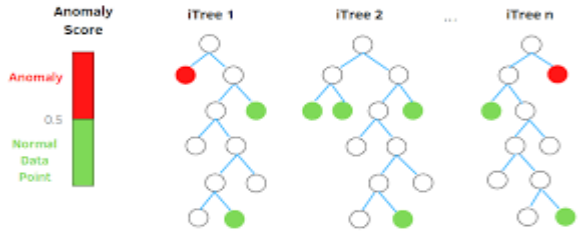

Fig-8 Isolation Forest Algorithm splitting

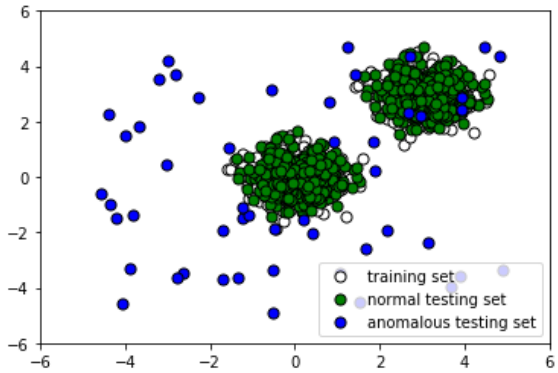

Fig-9 Scatter plot for Isolation Forest Algorithm

These results along with the classification report for isolation forest algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives.

```
Isolation Forest
Number of Errors: 71
Accuracy Score: 0.99750711000316

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.28      0.29      0.28        49

    accuracy                           1.00     28481
   macro avg       0.64      0.64      0.64     28481
weighted avg       1.00      1.00      1.00     28481
```

Fig-10 Results using 10% of the dataset

```
Isolation Forest
Number of Errors: 659
Accuracy Score: 0.9976861523768727

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.33      0.33      0.33       492

    accuracy                           1.00    284807
   macro avg       0.66      0.67      0.66    284807
weighted avg       1.00      1.00      1.00    284807
```

Fig-11 Results using the complete dataset

4. IMPLEMENTATION

Banks weren't willing to share sensitive data of its customers due to competition in market as well as some legal reasons. Hence, implementing this idea in real life was difficult. Now, a directive is issued by RBI that puts the responsibility to prove the consumer's fraud actions, on the banks. Earlier, customers were required to prove they're innocent. Bank customers have often declined to pay over disputes on transactions in their credit card bills. Hence, this proposed system can get bank's co-operation if it reduces the burden of the banks. We will apply the isolation forest algorithm on Amount class. When a user initiates a transaction, the requested amount is added to the existing dataset and then the algorithm is applied to the whole amount column. A score based on the results of the algorithm is assigned to each row and based on the score, anomalies are classified. In the given output, -1 means an anomaly and 1 is genuine transaction amount. Then all the anomalies in the amount column are displayed for visual comprehension.

5. RESULTS

If the initiated amount gets a score of 1, the transaction is approved and user gets a notification that his transaction was successful. If the anomaly score of initiated amounts is -1, the transaction is predicted to be a fraudulent one and the user gets an e-mail or a call from the bank to authorize the transaction. This adds an extra security layer to the existing system.

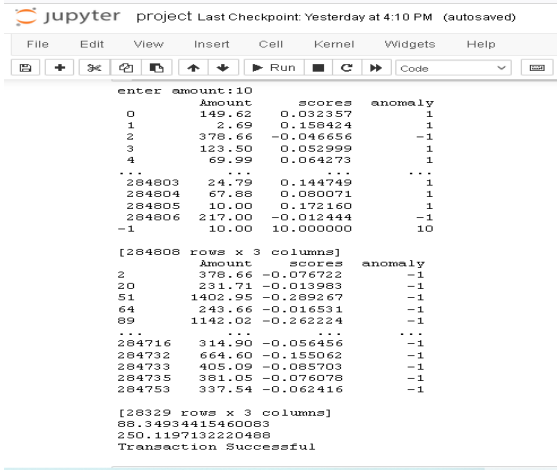

Fig-12 Successful transaction for a reasonable amount

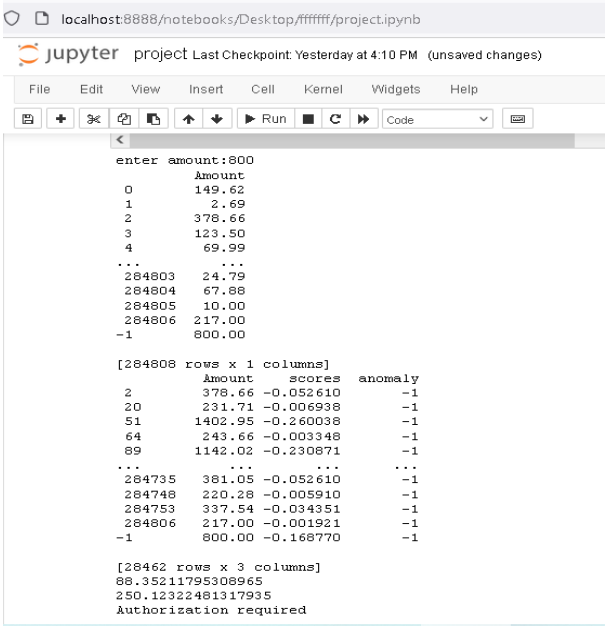

Fig-13 Authorization required for a predicted fraudulent amount

6. CONCLUSION

This paper aims to provide an insight on how machine learning can be used to predict fraud by banks and credit card companies to protect their users. Isolation Forest algorithm is used and it reaches over 99.7% accuracy but the precision is only 28% when a 10% of the data set is used. When the entire dataset is used in the algorithm, the precision increases to 33%. This high accuracy can be attributed to the huge imbalance between the number of valid and number of genuine transactions.

More machine learning algorithms can be integrated together to increase the efficiency of the system. Also, the algorithms can be applied to more columns like the time of transactions or the merchants to which the payment is being made to further predict the customer's behaviour more accurately. Since the entire dataset consists of only two days' transaction records, it's only a fraction of data that can be made available if this project were to be used on a commercial scale. The system is based on machine learning algorithms; hence it will increase its efficiency over time with addition of more data.

4. REFERENCES

[1] Clifton Phua1, Vincent Lee1, Kate Smith1 & Ross Gayler2 "A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australiae 1

[2] "Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence

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

[4] towardsdatascience.com for machine learning codes.

