CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

MINI PROJECT REPORT

Submitted by

SHASWATHA THILAKA.D (19EUEC135)

In partial fulfilment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING

in

ELECTRONICS AND COMMUNICATION ENGINEERING

SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY
(AN AUTONOMOUS INSTITUTION)
AFFILIATED TO ANNA UNIVERSITY CHENNAI
ACCREDITED BY NAAC WITH `A` GRADE

MARCH 2021

SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY

(An Autonomous Institution)

(Approved by AICTE and Affiliated to Anna University, Chennai)

ACCREDIATED BY NAAC WITH "A" GRADE

BONAFIDE CERTIFICATE

Certified that this project report **"CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING"** is the bonafide work of **SHASWATHA THILAKA.D (19EUEC35)** who carried out the project work under my supervision.

SIGNATURE

DR S SOPHIA Phd.,

HEAD OF THE DEPARTMENT

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY
KUNIYAMUTHUR
COIMBATORE-641008.

SIGNATURE

Ms G. SARANYA

ASSISTANT PROFESSOR

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY
KUNIYAMUTHUR
COIMBATORE-641008.

This project report submitted for the Autonomous Project Viva-voice examination held on_____

INTERNAL EXAMINER

EXTERNAL EXAMINER

ABSTRACT

Credit card plays a very important rule in today's economy. It becomes an unavoidable part of household, business and global activities. Although using credit cards provides enormous benefits when used carefully and responsibly, significant credit and financial damages may be caused by fraudulent activities. the classifier with better rating score can be chosen to be one of the best methods to predict frauds. Thus, followed by a feedback mechanism to solve the problem of concept. Many techniques have been proposed to confront thegrowthin credit card fraud. The main aim of the paper is to design and develop a novel fraud detection method for Streaming Transaction Data, with an objective, to analyse the past transaction details of the customers and extract the behavioural patterns. Where cardholders are clustered into different groups based on their transaction amount. The advantages and disadvantages of fraud detection methods are enumerated and compared.

Key teams: *Credit Card, Fraud Classification, Fraud Detection Techniques*

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In credit card transactions ‘Fraud’ is unauthorized and unwanted usage of an account by someone apart from the owner of that account. Due to rise and acceleration of E- Commerce, there has been a tremendous use of credit cards for online shopping which led to high amount of frauds associated to credit cards. Necessary prevention measures can be taken to prevent this abuse and therefore the behaviour of such fraudulent practices can be studied to reduce it and protect against similar occurrences in the future. In other words, Credit Card Fraud can be defined as a case where a person uses someone else’s credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used. In the era of digitalization, the necessity to identify credit card frauds is important. Fraud detection involves monitoring and analyzing the behaviour of various users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. This is a very relevant problem that demands the attention of communities like machine learning and data science where the answer to the problem are often automated. This is a really challenging problem from the perspective of learning, because it is characterized by various factors like class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time. In order to identify credit card fraud detection effectively, we need to understand the various

technologies, algorithms and types involved in detecting credit card frauds. Algorithm can differentiate transactions which are fraudulent or not. To find fraud, they need to pass dataset and knowledge of fraudulent transaction. They analyse the dataset and classify all transactions. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to verify if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is employed to coach and update the algorithm to eventually improve the fraud-detection performance over time. Here we model the sequence of operations in credit card transaction processing using a Hidden Markov Model (HMM) and show how it can be used for the detection of frauds. An HMM is initially trained with the normal behaviour of a cardholder. If an incoming credit card transaction is not accepted by the trained HMM with sufficiently high probability, it is considered to be fraudulent. At the same time, we try to ensure that genuine transactions are not rejected. We present experimental results to show the effectiveness of our approach and compare it with other techniques available in the literature.

1.1 Types of Fraud

The types of frauds considered in this paper are:

1. Credit Card Fraud: Credit card fraud is divided into two types:

i. Offline fraud: Offline fraud is done by using a stolen physical card at any place.

ii. On-line fraud: On-line fraud is committed over internet, phone, online shopping or when the card holder is not present.

2. Telecommunication Fraud - The use of telecommunication services to commit other forms of fraud. Consumers, businesses and communication service provider are the victims.

3. Computer Intrusion - Intrusion is defined as the act of entering without warrant or invitation; that means “potential possibility of unauthorized attempt to access Information, Manipulate Information Purposefully. Intruders may be from any environment, an outsider (Or Hacker) and an insider who knows the layout of the system.

4. Bankruptcy Fraud - Bankruptcy fraud means using a credit card while being absent. Bankruptcy fraud is one of the most complicated types of fraud to predict.

5. Theft Fraud/ Counterfeit Fraud - In this section, the focus is on theft and counterfeit fraud, which are related to one other. Theft fraud refers to the other person who is not the owner of the card. As soon as the owner give some feedback and contact the bank, the bank will take measures to check the thief as early as possible. Likewise, counterfeit fraud occurs when the credit card is used remotely; where only the credit card details are needed.

6. Application Fraud - When any people apply for a credit card with false information then it is termed as application fraud. For detecting application fraud, two different situations have to be classified. When applications come from a same user with the same details, that is called duplicates, and when applications come from different individuals with similar details, that is termed as identity fraudsters.

7. Internal Fraud - Banking sector allows their employees to access customer data. The data is the same information needed to access online banking to customer accounts. So the fraud can be done easily by an employee. Instead of this, financial institutions should require a password or PIN for net banking, and the password or PIN should be stored in the format of encrypted.

CHAPTER 2
LITERATURE SURVEY

Real-time Credit Card Fraud Detection Using Machine Learning

Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi– 2019

Credit card fraud events take place frequently and then result in huge financial losses. The number of online transactions has grown in large quantities and online credit card transactions holds a huge share of these transactions. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand. Fraudulent transactions can occur in various ways and can be put into different categories. This paper focuses on four main fraud occasions in real-world transactions. Each fraud is addressed using a series of machine learning models and the best method is selected via an evaluation. This evaluation provides a comprehensive guide to selecting an optimal algorithm with respect to the type of the frauds and we illustrate the evaluation with an appropriate performance measure. Another major key area that is addressed in this project is real-time credit card fraud detection. For this, the use of predictive analytics done by the implemented machine learning models and an API module to decide if a particular transaction is genuine or fraudulent. Also it assesses a novel strategy that effectively addresses the skewed distribution of data. The data used in the experiments come from a financial institution according to a confidential disclosure agreement.

Survey on Credit Card Fraud Detection Techniques

P. Jayant, Vaishali , D. Sharma Amity Institute of Information Technology Amity University, U.P

Due to a rapid advancement in the electronic commerce technology, the use of credit cards has dramatically increased. Since credit card is the most popular mode of payment, the number of fraud cases associated with it is also rising. In this paper, the survey on the present techniques available for detecting fraud in credit card is presented as a review paper. Fraud detection involves identifying fraud as quickly as possible once it has been done. Fraud detection methods are continuously developed to defend criminals in adapting to their strategies. The transaction is classified as normal, abnormal or suspicious depending on this initial belief. Once a transaction is found to be suspicious, belief is further strengthened or weakened according to its similarity with fraudulent or genuine transaction history using Bayesian learning.

A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective

SamanehSorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi Department of Information Technology, University of Guilan , Iran.

Credit card plays a very important rule in today's economy. It becomes an unavoidable part of household, business and global activities. Although using

credit cards provides enormous benefits when used carefully and responsibly, significant credit and financial damages may be caused by fraudulent activities. Many techniques have been proposed to confront the growth in credit card fraud. However, all of these techniques have the same goal of avoiding the credit card fraud; each one has its own drawbacks, advantages and characteristics. In this paper, after investigating difficulties of credit card fraud detection, we seek to review the state of the art in credit card fraud detection techniques, datasets and evaluation criteria. Furthermore, a classification of mentioned techniques into two main fraud detection approaches, namely, misuses (supervised) and anomaly detection (unsupervised) is presented. Again, a classification of techniques is proposed based on capability to process the numerical and categorical datasets. Different datasets used in literature are then described and grouped into real and synthesized data and the effective and common attributes are extracted for further usage. Moreover, evaluation employed criterions in literature are collected and discussed. Consequently, open issues for credit card fraud detection are explained as guidelines for new researchers.

Credit Card Fraud Detection Techniques: A Review

Sonal Mehndiratta , Mr. Kamal Gupta Hod and Assistant professor Guru Nanak Institute of Technology Mullana, Ambala .

The prediction analysis is the approach which can predict future possibilities on the current data. When the physical-card based purchasing technique is applied, the card is given by the cardholder to the merchant so that a successful payment method can be performed. The fraudulent transactions are conducted by the attacker by stealing the credit card. When the loss of the card is not noticed by the cardholder, a huge loss can be faced by the credit card company. A very little amount of information is required by the attacker for conducting any fraudulent transaction in online transactions. In this research work, various credit card fraud detection techniques are reviewed in terms of certain parameters.

CHAPTER 3
EXISTING SYSTEM

Credit-card-based purchases can be categorized into two types:

1) Physical card purchase and

2) Virtual card purchase.

In a physical card purchase, the cardholder personally presents the card to make a payment. While doing a physical card purchase, the attacker needs to steal the credit card and forge the signature in order to make a purchase. In the virtual card purchase, only the card information is required such as card number, expiration date, secure code, etc. Such purchases are normally done on the Internet or over telephone. To commit fraud in these types of purchases, a person simply needs to know the card details. The mode of payment for online purchase is mostly done by credit card. The amount of financial losses due to credit card frauds is growing as the usage of the credit cards is common. Security means to use credit card safely and avoid the occurrence of fraud. The purpose of security is to avoid fraudulent usage of credit cards. In Fraud cases issues like lost cards, stolen lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud and non-received issue (NRI) fraud are found. For decreasing these frauds, security with credit cards is needed.

3.1 Types of Solution for detection of fraud

Usually fraud is discovered when a credit card is lost or stolen, when unfamiliar charges on the billing statement are found, when calls or letters about transactions that have not been made, contacted by the credit card company's fraud department to question about the charge. If the fraud is suspected on the account, then one should contact the credit card company immediately. The credit card company will be able to help in verifying the fraud, remove the charges which have not been used by the card holder or any authorized person, close down the account to prevent more fraudulent transactions and issue a new account number and new card, and transfer old information to the new account. It's also a good idea to check credit report to be sure there's nothing else that looks suspicious. In most cases, the involvement of law enforcement will be coordinated with the financial institution.

Identity theft is a particular type of fraud in which a thief uses the personal information to set up new accounts or get other benefits in the name of cardholder. Though it's not as common as other types of fraud, it can be more challenging and cause more severe problems. Some signs of identity theft are: cardholder is not receiving the bills or other mail, receives credit card, being denied credit for no apparent reason, getting calls or letters about things that were not transacted by credit cardholder, being served court papers or arrest warrants for things in which there is no involvement of cardholder.

Different techniques have been applied to detect the frauds that occur in credit card transactions. These techniques are explained below:

a. Artificial Neural Network: A set of interlinked nodes that are designed for imitating the working of a human brain is known as an artificial neural network (ANN). A weighted link is assigned to all the other nodes that are present in the adjacent layers of each node.

b. Genetic Algorithm (GA): The genetic algorithms were introduced inspiring from natural evolution. Chromosomes are the binary strings that are used to represent the populations of candidate solutions. It is based on the concept that the chances of survival and reproduction are higher for the chromosomes with higher quality i.e. having better fitness value.

c. Hidden Markov Model (HMM): A double embedded stochastic process using which highly complicated stochastic processes can be generated is known as a hidden Markov model. A Markov process that has unobserved states is assumed to be available within the underlying system. The only unknown parameters are the definite transition of the states within the simpler Markov models

d. KNN Classifier: KNN is the non-parametric algorithm used in case of classification and regression. In classification and regression, the input is consisting of K-nearest training examples in the feature space and on the other hand, the output depends upon whether KNN belongs to regression category or classification category.

e. Naïve Bayes: This algorithm implements the Bayesian rule on categorical data for performing classification on it. In comparison to other classification approaches, the performance of the Naïve Bayes algorithm is known to be better and very simple

CHAPTER-4

PROPOSED SYSTEM

4.1 PROPOSED TECHNIQUIE:

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers. First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data. The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one. After checking this dataset, we plot a histogram for every column. This is done to get a graphical representation of the dataset which can be used to verify that there are no missing any values in the dataset. This is done to ensure that we don't require any missing value imputation and the machine learning algorithms can process the dataset smoothly. After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable. The dataset is now formatted and processed. The time and amount column are standardized and the Class column is removed to ensure fairness of evaluation. The data is processed by a set of algorithms

from modules. The following module diagram explains how these algorithms work together:

This data is fit into a model and the following outlier detection modules are applied on it:

• Local Outlier Factor

• Isolation Forest Algorithm

These algorithms are a part of sklearn. The ensemble module in the sklearn package includes ensemble-based methods and functions for the classification, regression and outlier detection. Thisand machine learning. It features various classification, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries. We've used Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests. This program can also be executed on the cloud using Google Collab platform which supports all python notebook files. free and open-source Python library is built using NumPy, SciPy and matplotlib modules which provides a lot of simple and efficient tools which can be used for data analysis.

4.2 ADVANTAGES

1. Higher accuracy of fraud detection. Compared to rule-based solutions, machine learning tools have higher precision and return more relevant results as they consider multiple additional factors. This is because ML technologies can consider many more data points, including the tiniest details of behaviour patterns associated with a particular account.

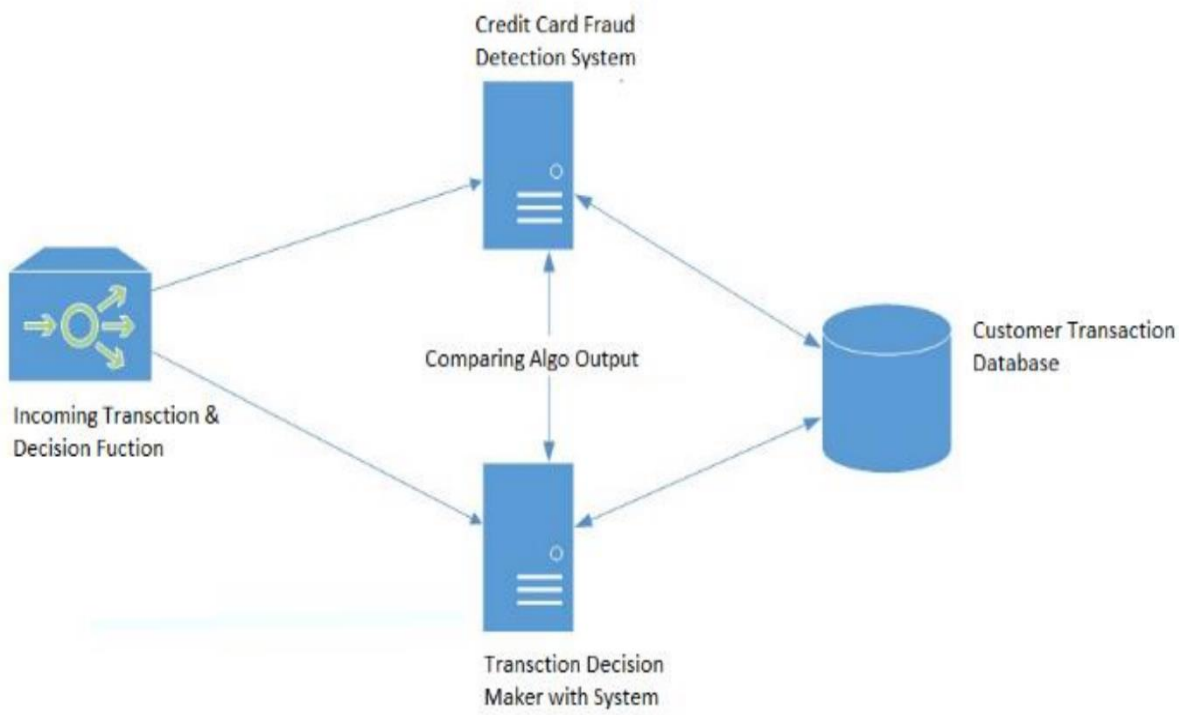
2. Less manual work needed for additional verification. Enhanced accuracy leads reduce the burden on analysts.

3. Fewer false declines. False declines or false positives happen when a system identifies a legitimate transaction as suspicious and wrongly cancels it.

4. Ability to identify new patterns and adapt to changes. Unlike rule-based systems, ML algorithms are aligned with a constantly changing environment and financial conditions. They enable analysts to identify new suspicious patterns and create new rules to prevent new types of scams.

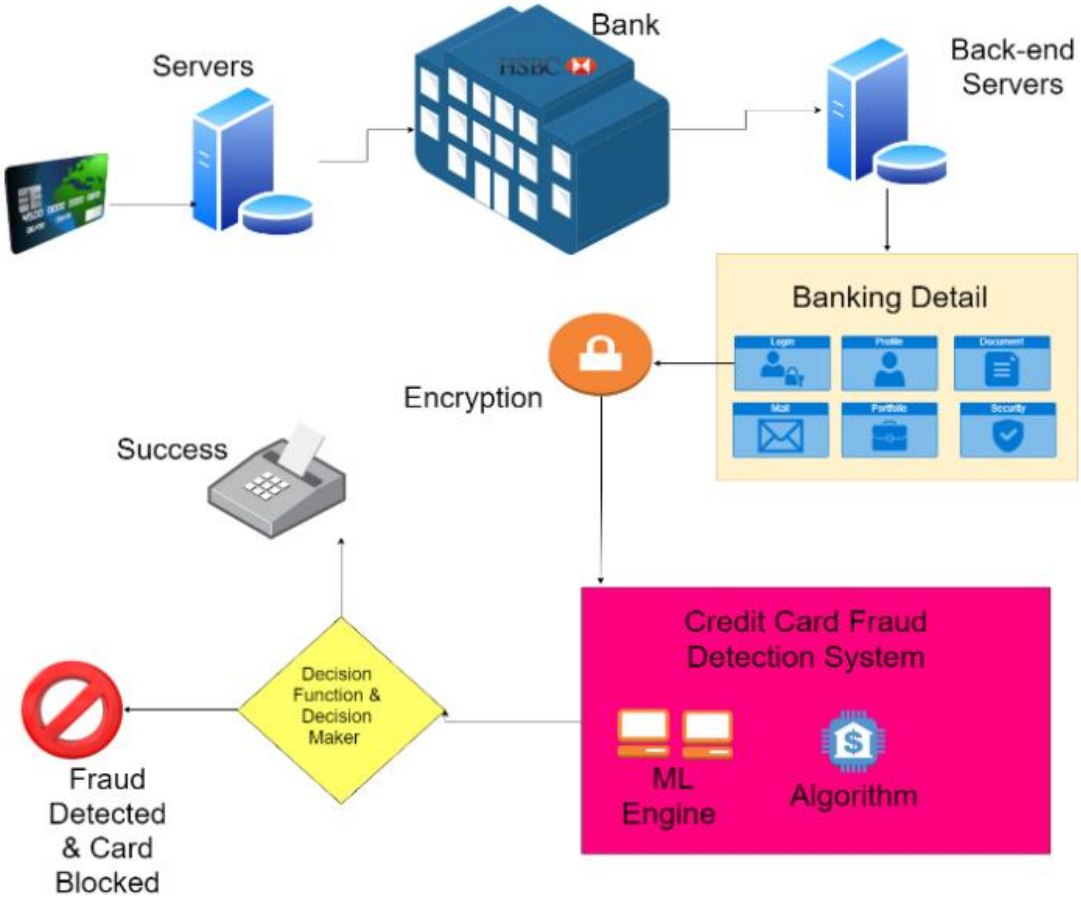
4.3 BLOCK DIAGRAM

The basic rough architecture diagram can be represented with the following figure:



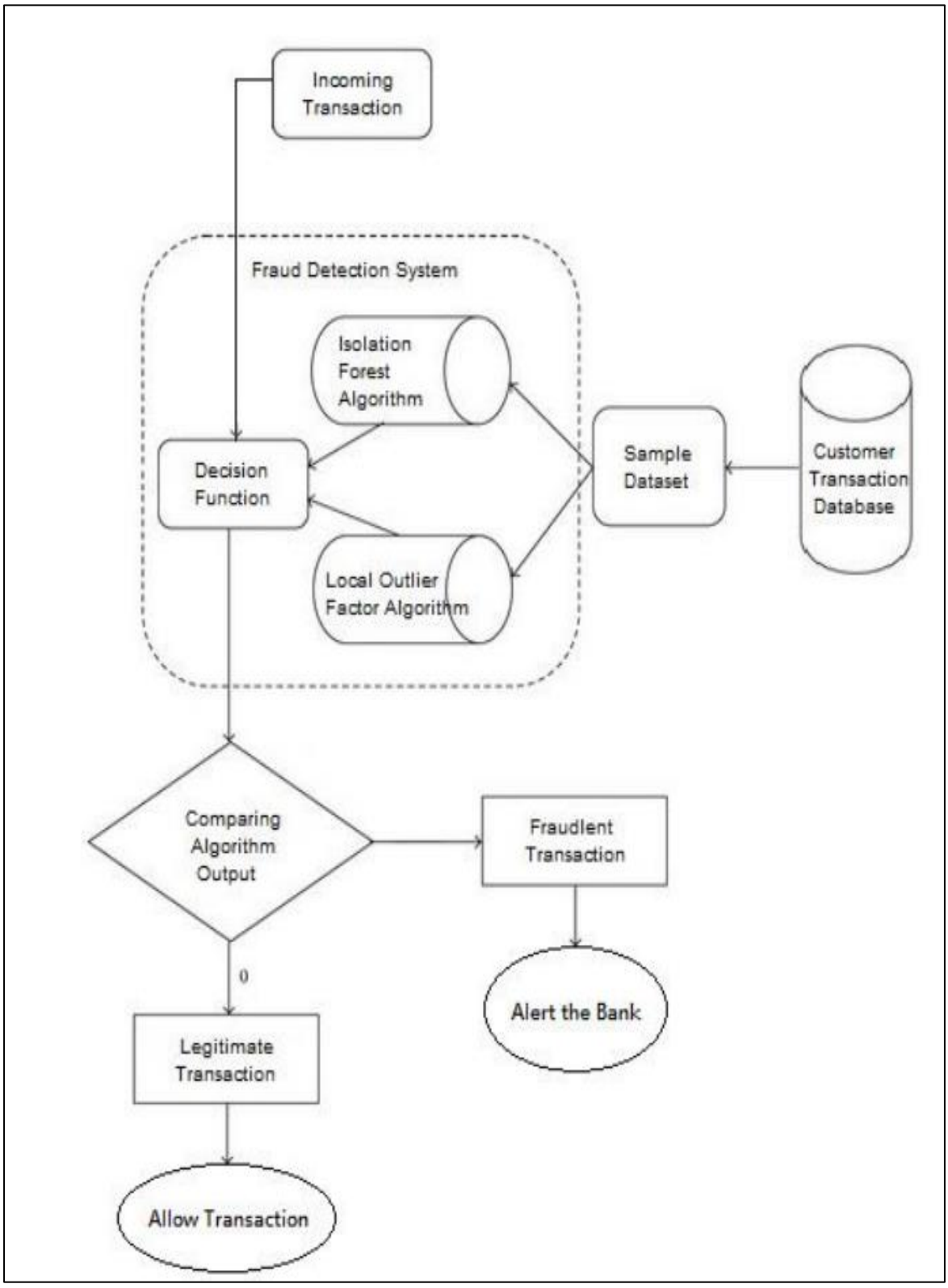
4.1 Basic Block Diagram of credit card fraud detection

When looked at in detail on a larger scale along with real life elements, the full architecture diagram can be represented as follows:



4.2 Architecture of credit card fraud detection

4.4 FLOWCHART



4.3 Flow chart of credit card fraud detection

CHAPTER 5

SOFTWARE DESCRIPTION

5.1 PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985-1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This tutorial gives enough understanding on Python programming language.

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- Python is Interactive − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

- Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

5.2 JUPYTER NOTEBOOK

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

The Jupyter Notebook is quite useful not only for learning and teaching a programming language such as Python but also for sharing your data.

5.1 Jupyter Notebook

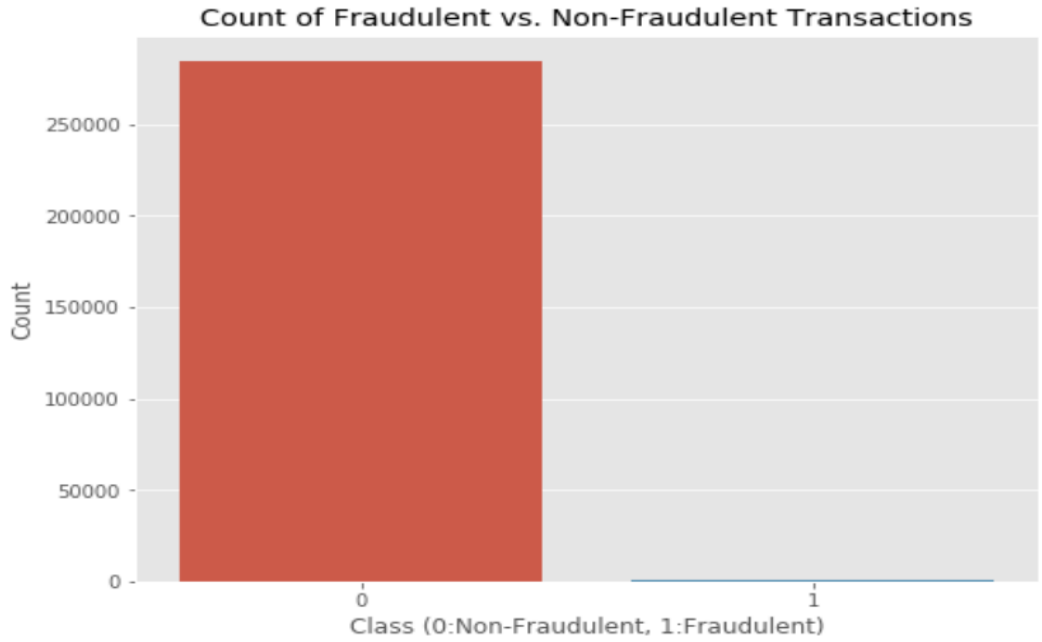
You can turn your Notebook into a slideshow or share it online with GitHub. If you want to share a Notebook without requiring your users to install anything, you can use binder for that.

Google and Microsoft both have their own version of the Notebook that you can use to create and share your Notebooks at Google Colaboratory and Microsoft Azure Notebooks respectively. You can browse really interesting Notebooks there as well.

Project Jupyter recently launched their latest product, JupyterLab. JupyterLab incorporates Jupyter Notebook into an Integrated Development type Editor that you run in your browser. You can kind of think of JupyterLab as an advanced version of Jupyter Notebook. JupyterLab allows you to run terminals, text editors and code consoles in your browser in addition to Notebooks.

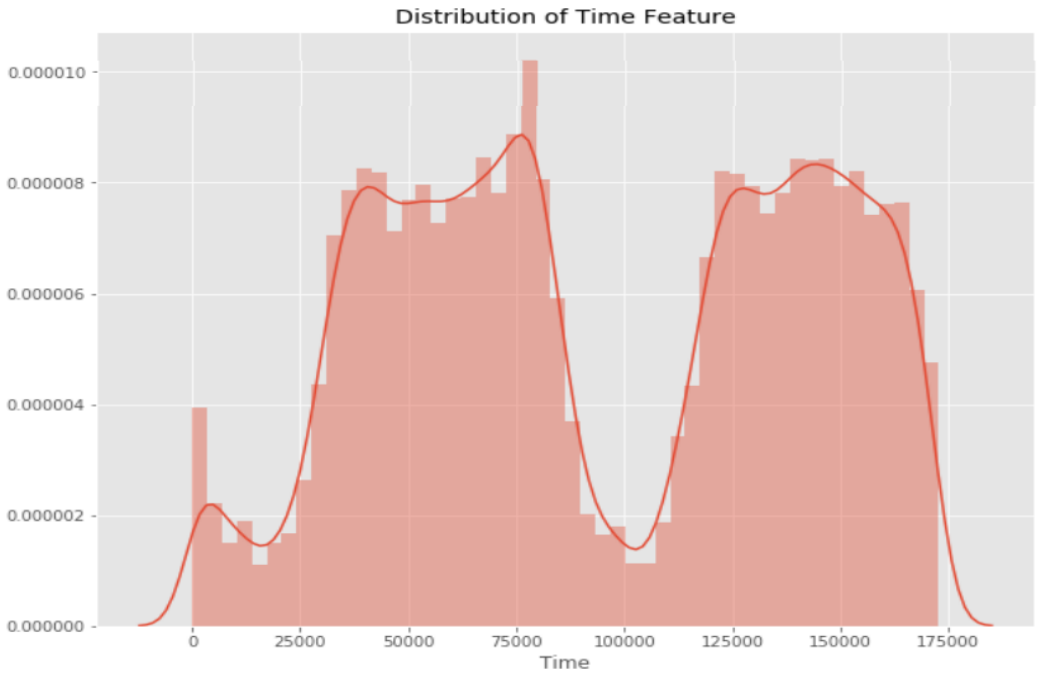
5.3 JUPITER NOTEBOOK IMPLEMENTATION

We plot different graphs to check for inconsistencies in the dataset and to visually comprehend it:



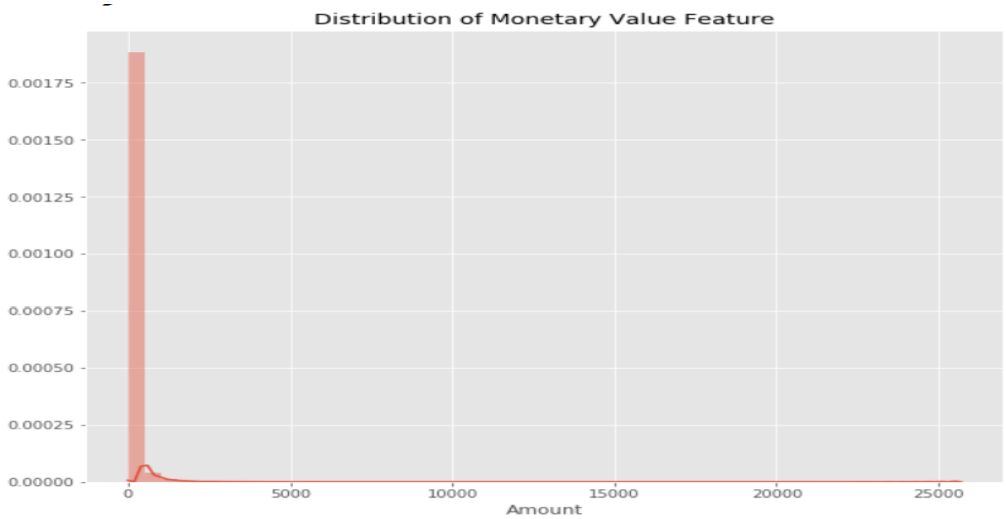
5.2 Fraudulent vs Non-Fraudulent Transaction

This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.



5.3 Distribution of Time feature

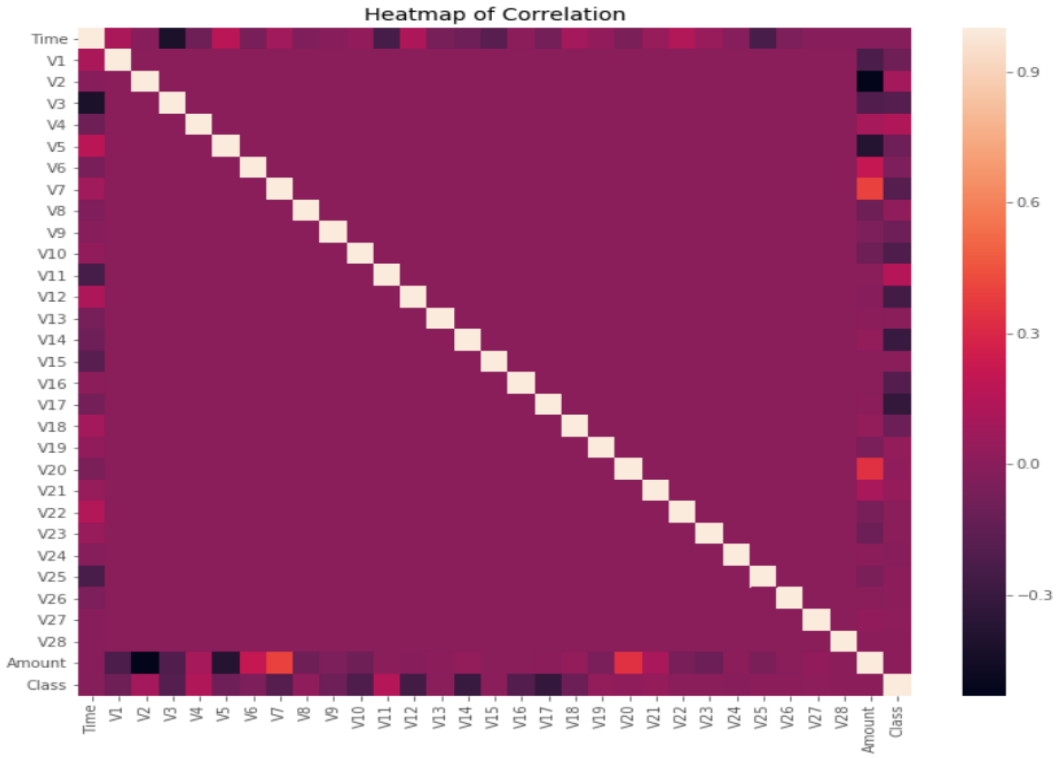
This graph shows the times at which transactions were done within two days. It can be seen that the least number of transactions were made during night time and highest during the days.



5.4 Distribution of Monetary value feature

This graph represents the amount that was transacted. A majority of transactions are relatively small and only a handful of them come close to the maximum transacted amount.

After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable. This heatmap is shown below:



5.5 Heatmap of correlation

A. Local Outlier Factor It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbours. More

precisely, locality is given by k-nearest neighbours, whose distance is used to estimate the local data. The pseudocode for this algorithm is written as:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

rng = np.random.RandomState(42)

# Generate train data
X = 0.3 * rng.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * rng.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = rng.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = IsolationForest(behaviour='new', max_samples=100,
                      random_state=rng, contamination='auto')
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)

# plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

5.6 pseudocode for local data

By comparing the local values of a sample to that of its neighbours, one can identify samples that are substantially lower than their neighbours. These values are quite amanous and they are considered as outliers. As the dataset is very large, we used only a fraction of it in out tests to reduce processing times. The final result with the complete dataset processed is also determined and is given in the results section of this paper.

B. Isolation Forest Algorithm The Isolation Forest ‘isolates’ observations by arbitrarily selecting a feature and then randomly selecting a split value between

the maximum and minimum values of the designated feature. Recursive partitioning can be represented by a tree, the number of splits required to isolate a sample is equivalent to the path length root node to terminating node. The average of this path length gives a measure of normality and the decision function which we use. The pseudocode for this algorithm can be written as:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

np.random.seed(42)

# Generate train data
X = 0.3 * np.random.randn(100, 2)
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
X = np.r_[X + 2, X - 2, X_outliers]

# fit the model
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[200:]

# plot the level sets of the decision function
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf._decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

5.7 Pseudocode of Isolation Forest Algorithm

Partitioning them randomly produces shorter paths for anomalies. When a forest of random trees mutually produces shorter path lengths for specific samples, they are extremely likely to be anomalies. Once the anomalies are detected, the system can be used to report them to the concerned authorities. For testing purposes, we are comparing the outputs of these algorithms to determine their accuracy and precision.

This idea is difficult to implement in real life because it requires the cooperation from banks, which aren't willing to share information due to their market competition, and also due to legal reasons and protection of data of their users. Therefore, we looked up some reference papers which followed similar approaches and gathered results. As stated in one of these reference papers: "This technique was applied to a full application data set supplied by a German bank in 2006. For banking confidentiality reasons, only a summary of the results obtained is presented below.

CHAPTER 6

RESULT

The code prints out the number of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end and both the results are printed. These results along with the classification report for each 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. Results when 10% of the dataset is used:

```
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


Local Outlier Factor
Number of Errors: 97
Accuracy Score: 0.9965942207085425

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

6.1 Results when 10% of the dataset

Results with the complete dataset is used:

```
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

Local Outlier Factor
Number of Errors: 935
Accuracy Score: 0.9967170750718908

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.05      0.05      0.05       492

    accuracy                           1.00    284807
   macro avg       0.52      0.52      0.52    284807
weighted avg       1.00      1.00      1.00    284807
```

6.2 Results with the complete dataset

CHAPTER 7

CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results. While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions. Since the entire dataset consists of only two days' transaction records, its only a fraction of data that can be made available if this project were to be used on a commercial scale. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

CHAPTER 8

FUTURE ENHANCEMENT

While we couldn’t reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

CHAPTER 9

REFERENCE

[1] "Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2] 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, Australia

[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] "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

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

[6] Credit Card Fraud Detection using Machine Learning and Data Science S P Maniraj Assistant Professor (O.G.) Department of Computer Science and Engineering SRM Institute of Science and Technology Aditya Saini, Swarna Deep Sarkar Shadab Ahmed Department of Computer Science and Engineering SRM Institute of Science and Technology.

[7] Real-time Credit Card Fraud Detection Using Machine Learning Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi– 2019