****

**DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY**

**Project Title:**

**CREDIT CARD FRAUD DETECTION SYSTEM**

**OMAR KHALID OMAR**

**C026-01-0300/2018**

**A project submitted to the Department of Computer Science in the School of Computer science and IT in partial fulfillment for the award of Bachelor of Science Degree in Computer Science.**

**2020.**

**DECLARATION**

This proposal is my original work and has not been presented for a degree in any other University

Name: ………………………………………………………………………………………..

……………………….…. ………………………………

Signature Date

This proposal has been submitted for examination with my approval as University Supervisor

Name: ………………………………………………………………………………………..

……………………… ……………………………………

Signature Date

**ABSTRACT**

It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications. Credit Card Fraud Detection is a typical sample of classification. In this process, we have focused on analyzing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms on the Credit Card Transaction dataset.

Contents

[Table of figures vi](#_Toc33112749)

[CHAPTER ONE: INTRODUCTION 1](#_Toc33112750)

[1.1 Background Information 1](#_Toc33112751)

[1.2 PROBLEM STATEMENT 1](#_Toc33112752)

[1.3 System objectives 2](#_Toc33112753)

[1.3.1 General objectives 2](#_Toc33112754)

[1.3.2 Specific Objectives 2](#_Toc33112755)

[1.4 Scope 3](#_Toc33112756)

[1.4.1 Target Users. 3](#_Toc33112757)

[1.4.2 Platform and where to implement**.** 3](#_Toc33112758)

[CHAPTER TWO: LITERATURE REVIEW 4](#_Toc33112759)

[2.0 Introduction 4](#_Toc33112760)

[2.1 EXISTING SYSTEMS 5](#_Toc33112761)

[2.1.0 Fraud Prevention Technologies 5](#_Toc33112762)

[2.1.1 Recent Developments in Fraud Management 6](#_Toc33112763)

[CHAPTER THREE: METHODOLOGY 10](#_Toc33112764)

[3.1 INTRODUCTION 10](#_Toc33112765)

[3.2 Data Collection 10](#_Toc33112766)

[3.3 Data Analysis 10](#_Toc33112767)

[3.3.1 Pie chart of the dataset 11](#_Toc33112768)

[3.3.2 Transactions distribution of original dataset 12](#_Toc33112769)

[3.3.3 Random over-sampling balancing technique output 13](#_Toc33112770)

[3.3.4 Random under-sampling balancing technique output 14](#_Toc33112771)

[3.3.5 Synthetic Minority Over-sampling Technique (SMOTE) balancing output 15](#_Toc33112772)

[3.4 Tools 15](#_Toc33112773)

[CHAPTER FOUR: SYSTEM MODELLING 16](#_Toc33112774)

[4.1 Introduction 16](#_Toc33112775)

[4.2 Requirements Analysis 16](#_Toc33112776)

[4.2.1 Functional Requirements 16](#_Toc33112777)

[4.2.2 Non Functional Requirements 16](#_Toc33112778)

[4.4 System Analysis 17](#_Toc33112779)

[4.4.1 Use case diagram 17](#_Toc33112780)

[4.4.2 Sequence Diagram 18](#_Toc33112781)

[4.4.3 Class diagram 18](#_Toc33112782)

[4.4.4 Data flow diagram 19](#_Toc33112783)

[CHAPTER FIVE: SYSTEM IMPLEMENTATION 20](#_Toc33112784)

[5.0 Introduction 20](#_Toc33112785)

[5.1 Tools used for coding 20](#_Toc33112786)

[5.2 System testing and Implementation 20](#_Toc33112787)

[5.2.1 Objectives 20](#_Toc33112788)

[5.2.2 Testing Overview 20](#_Toc33112789)

[5.2.3 Testing strategy 20](#_Toc33112790)

[5.2.4 Test plan 20](#_Toc33112791)

[5.3 Implementation Plan 21](#_Toc33112792)

[5.3.1 Process in Implementation plan 21](#_Toc33112793)

[CHAPTER SIX: LIMITATION, CONCLUSION AND RECOMMENDATION. 22](#_Toc33112794)

[6.0 Introduction 22](#_Toc33112795)

[6.1 Limitations from the project 22](#_Toc33112796)

[6.2 Conclusion 22](#_Toc33112797)

[6.3 Recommendations 22](#_Toc33112798)

[References 24](#_Toc33112799)

# Table of figures

[Figure 1: Pie chart for credit card transactions 11](#_Toc509568923)

[Figure 2:Scatter plot for credit card transactions original dataset 12](#_Toc509568924)

[Figure 3: Scatter plot for ROS output 13](#_Toc509568925)

[Figure 4: Scatter plot for RUS output 14](#_Toc509568926)

[Figure 5: Scatter plot for SMOTE output 15](#_Toc509568927)

[Figure 6: Use case diagram 17](#_Toc509568928)

[Figure 7: Sequence diagram 18](#_Toc509568929)

[Figure 8: Class diagram 18](#_Toc509568930)

[Figure 9: Data flow diagram 19](#_Toc509568931)

# CHAPTER ONE: INTRODUCTION

## 1.1 Background Information

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behavior of such fraudulent practices can be studied to minimize 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. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behavior, which consist of fraud, intrusion, and defaulting. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated

## 1.2 PROBLEM STATEMENT

This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as 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. These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyze all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time. Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies.

These frauds are classified as:

* Credit Card Frauds: Online and Offline
* Card Theft
* Account Bankruptcy
* Device Intrusion
* Application Fraud
* Counterfeit Card
* Telecommunication Fraud

Some of the currently used approaches to detection of such fraud are:

* Artificial Neural Network
* Fuzzy logic
* Genetic Algorithm
* Logistic Regression
* Decision Tree
* Support Vector Machines
* Bayesian Networks
* Hidden Markov Model
* K-Nearest Neighbor

## 1.3 System objectives

### 1.3.1 General objectives

The objectives of credit card fraud detection are to reduce losses due to payment fraud for both merchants and issuing banks and increase revenue opportunities for merchants.

### 1.3.2 Specific Objectives

1. To prevent unauthorized and unwanted usage of an account by someone other than the owner of that account.
2. To come up with a system that views all transactions of a cardholder and authorize non-fraudulent transactions while reporting suspicious ones to the investigators.
3. To come up with a system that learns the patterns of each cardholder’s transactions to improve performance over time.
4. To minimize economic loss to banks and financial institutions as well as cardholders. This fraud affects up to the country’s economy.

## 1.4 Scope

This section explains the scope of the system including the target users, where to implement and specific platform used.

### 1.4.1 Target Users.

The main user of the system is banks and financial institutions.

### 1.4.2 Platform and where to implement**.**

The system is developed by R language on RStudio IDE and it will be implemented by investigators of banks, financial institutions and even security personnel.

# CHAPTER TWO: LITERATURE REVIEW

## 2.0 Introduction

Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule or policy with an aim to attain unauthorized financial benefit. Numerous literatures pertaining to anomaly or fraud detection in this domain have been published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates have revealed that techniques employed in this domain include data mining applications, automated fraud detection, adversarial detection. In another paper, Suman, Research Scholar, GJUS&T at Hisar HCE presented techniques like Supervised and Unsupervised Learning for credit card fraud detection. Even though these methods and algorithms fetched an unexpected success in some areas, they failed to provide a permanent and consistent solution to fraud detection. A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of credit card transaction data set of one certain commercial bank. Outlier mining is a field of data mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e. the transactions that aren’t genuine. They have taken attributes of customer’s behavior and based on the value of those attributes they’ve calculated that distance between the observed value of that attribute and its predetermined value. Unconventional techniques such as hybrid data mining/complex network classification algorithm is able to perceive illegal instances in an actual card transaction data set, based on network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group have proved efficient typically on medium sized online transaction. There have also been efforts to progress from a completely new aspect. Attempts have been made to improve the alert feedback interaction in case of fraudulent transaction. In case of fraudulent transaction, the authorized system would be alerted and a feedback would be sent to deny the ongoing transaction. Artificial Genetic Algorithm, one of the approaches that shed new light in this domain, countered fraud from a different direction. It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

## 2.1 EXISTING SYSTEMS

### 2.1.0 Fraud Prevention Technologies

While fraudsters are using sophisticated methods to gain access to credit card information and perpetrate fraud, new technologies are available to help merchants to detect and prevent fraudulent transactions. Fraud detection technologies enable merchants and banks to perform highly automated and sophisticated screenings of incoming transactions and flagging suspicious transactions. The various fraud prevention techniques are discussed below:

**Manual review**

This method consists of reviewing every transaction manually for signs of fraudulent activity and involves an exceedingly high level of human intervention. This can prove to be very expensive, as well as time consuming. Moreover, manual review is unable to detect some of the more prevalent patterns of fraud, such as use of a single credit card multiple times on multiple locations (physical or web sites) in a short span.

**Card verification methods**

The Card Verification Method3 (CVM) consists of a 3- or 4-digit numeric code printed on the card but is not embossed on the card and is not available in the magnetic stripe. The merchant can request the cardholder to provide this numeric code in case of card-not-present transaction and submit it with authorization. The purpose of CVM is to ensure that the person submitting the transaction is in possession of the actual card, since the code cannot be copied from receipts or 55 skimmed from magnetic stripe. Although CVM provides some protection for the merchant, it doesn’t protect them from transactions placed on physically stolen cards. Furthermore, fraudsters who have temporary possession of a card could, in principle, read and copy the CVM code.

**Negative and Positive lists**

A negative list is a database used to identify high-risk transactions based on specific data fields. An example of a negative list would be a file containing all the card numbers that have produced charge-backs in the past, used to avoid further fraud from repeat offenders. Similarly a merchant can build negative lists based on billing names, street addresses, emails and internet protocols (IPs) that have resulted in fraud or attempted fraud, effectively blocking any further attempts. A merchant/acquirer could create and maintain a list of high-risk countries and decide to review or restrict orders originating from those countries. Another popular example of negative list is the SAFE file distributed by MasterCard to merchants and member banks. This list contains card numbers, which could be potentially used by fraudsters, e.g., cards that have been reported as lost or stolen in the immediate recent past. Positive files are typically used to recognize trusted customers, perhaps by their card number or email address, and therefore bypass certain checks. Positive files represent an important tool to prevent unnecessary delays in processing valid orders.

**Payer authentication**

Payer authentication is an emerging technology that promises to bring in a new level of security to business-to-consumer internet commerce. The first 56 implementation of this type of service is the Verified by Visa (VbV) or Visa Payer Authentication Service (VPAS) program, launched worldwide by Visa in 2002. The program is based on a Personal Identification Number (PIN) associated with the card, similar to those used with ATM cards, and a secure direct authentication channel between the consumer and the issuing bank. The PIN is issued by the bank when the cardholder enrolls the card with the program and will be used exclusively to authorize online transactions. When registered cardholders check out at a participating merchant’s site, they will be prompted by their issuing bank to provide their password. Once the password is verified, the merchant may complete the transaction and send the verification information on to their acquirer.

**Lockout mechanisms**

Automatic card number generators represent one of the new technological tools frequently utilized by fraudsters. These programs, easily downloadable from the Web, are able to generate thousands of ‘valid’ credit card numbers. The traits of frauds initiated by a card number generator are the following: • Multiple transactions with similar card numbers (e.g. same Bank Identification Number (BIN)) • A large number of declines Acquiring banks/merchant sites can put in place prevention mechanisms specifically designed to detect number generator attacks.

**Fraudulent merchants**

Both MasterCard and Visa publish a list of merchants who have been known for being involved in fraudulent transactions in the past. These lists (NMAS - 57 from Visa and MATCH - from MasterCard) could provide useful information to acquirers’ right at the time of merchant recruitment preventing potential fraudulent transactions.

# 2.1.1 Recent Developments in Fraud Management

The technology for detecting credit card frauds is advancing at a rapid pace – rules based systems, neural networks, chip cards and biometrics are some of the popular techniques employed by Issuing and Acquiring banks these days. Apart from technological advances, another trend which has emerged during the recent years is that fraud prevention is moving from back-office transaction processing systems to front-office authorization systems to prevent committing of potentially fraudulent transactions. However, this is a challenging trade-off between the response time for processing an authorization request and extent of screening that should be carried out.

**Simple rule systems**

Simple rule systems involve the creation of ‘if...then’ criteria to filter incoming authorizations/transactions. Rule-based systems rely on a set of expert rules designed to identify specific types of high-risk transactions. Rules are created using the knowledge of what characterizes fraudulent transactions. For instance, a rule could look like – If transaction amount is > $5000 and card acceptance location = Casino and Country = ‘a high-risk country’. Fraud rules enable to automate the screening processes leveraging the knowledge gained over time regarding the characteristics of both fraudulent and legitimate transactions. Typically, the effectiveness of a rule-based system will increase over time, as more rules are added 58 to the system. It should be clear, however, that ultimately the effectiveness of the system depends on the knowledge and expertise of the person designing the rules. The disadvantage of this solution is that it can increase the probability of throwing many valid transactions as exceptions; however, there are ways by which this limitation can be overcome to some extent by prioritizing the rules and fixing limits on number of filtered transactions.

**Risk scoring technologies**

Risk scoring tools are based on statistical models designed to recognize fraudulent transactions, based on a number of indicators derived from the transaction characteristics. Typically, these tools generate a numeric score indicating the likelihood of a transaction being fraudulent: the higher the score, the more suspicious the order. Risk scoring systems provide one of the most effective fraud prevention tools available. The primary advantage of risk scoring is the comprehensive evaluation of a transaction being captured by a single number. While individual fraud rules typically evaluate a few simultaneous conditions, a risk-scoring system arrives at the final score by weighting several dozens of fraud indicators, derived from the current transaction attributes as well as cardholder historical activities. E.g., transaction amounts more than three times the average transaction amount for the cardholder in the last one year. The second advantage of risk scoring is that, while a fraud rule would either flag or not flag a transaction, the actual score indicates the degree of suspicion on each transaction. 59 Thus, transactions can be prioritized based on the risk score and given a limited capacity for manual review, only those with the highest score would be reviewed.

**Neural network technologies**

Neural networks are an extension of risk scoring techniques. They are based on the ‘statistical knowledge’ contained in extensive databases of historical transactions, and fraudulent ones in particular. These neural network models are basically ‘trained’ by using examples of both legitimate and fraudulent transactions and are able to correlate and weigh various fraud indicators (e.g., unusual transaction amount, card history, etc) to the occurrence of fraud. A neural network is a computerized system that sorts data logically by performing the following tasks:

• Identifies cardholder’s buying and fraudulent activity patterns.

• Processes data by trial and elimination (excluding data that is not relevant to the pattern).

• Finds relationships in the patterns and current transaction data. The principles of neural networking are motivated by the functions of the brain – especially pattern recognition and associative memory. The neural network recognizes similar patterns, predicting future values or events based upon the associative memory of the patterns it has learned. The advantages neural networks offer over other techniques are that these models are able to learn from the past and thus, improve results as time passes. They can also extract rules and predict future activity based on the current situation. By employing neural networks effectively, banks can detect fraudulent use of a card, faster and more efficiently.

**Biometrics**

Biometrics is the name given to a fraud prevention technique that records a unique characteristic of the cardholder like, a fingerprint or how he/she sign his/her name, so that it can be read by a computer. The computer can then compare the stored characteristic with that of the person presenting the card to make sure that the right person has the right card. Biometrics, which provides a means to identify an individual through the verification of unique physical or behavioral characteristics, seems to supersede PIN as a basis for the next generation of personal identity verification systems. There are many types of biometrics systems under development such as finger print verification, hand based verification, retinal and iris scanning and dynamic signature verification.

**Address verification system**

This technique is applicable in card-not-present scenarios. Address Verification System (AVS) matches the first few digits of the street address and the ZIP code information given for delivering/billing the purchase to the corresponding information on record with the card issuers. A code representing the level of match between these addresses is returned to the merchant. AVS is not much useful in case of international transactions.

**Smart cards**

To define in the simplest terms, a smart card is a credit card with some intelligence in the form of an embedded CPU. This card-computer can be programmed to perform tasks and store information, but the intelligence is limited – meaning that the smart card's power falls far short of a desktop computer. 61 Smart credit cards operate in the same way as their magnetic counterparts, the only difference being that an electronic chip is embedded in the card. These smart chips add extra security to the card. Smart credit cards contain 32-kilobyte microprocessors, which is capable of generating 72 quadrillion or more possible encryption keys and thus making it practically impossible to fraudulently decode information in the chip. The smart chip has made credit cards a lot more secure; however, the technology is still being run alongside the magnetic strip technology due to a slow uptake of smart card reading terminals in the world market. Smart cards have evolved significantly over the past decade and offer several advantages compared to a generalpurpose magnetic stripe card. The advantages are listed below:

• Stores many times more information than a magnetic stripe card.

• Reliable and harder to tamper with than a magnetic stripe card.

• Performs multiple functions in a wide range of industries.

• Compatible with portable electronic devices such as phones and personal digital assistants (PDAs), and with PCs.

• Stores highly sensitive data such as signing or encryption keys in a highly secure manner

• Performs certain sensitive operations using signing or encryption keys in a secure fashion.

A consortium of Europay MasterCard and Visa (EMV) recently issued a set of specifications for embedding chips in credit cards and processing transactions 62 from such cards. MasterCard and Visa have also issued deadlines for compliance with these specifications indicating that banks will have to bear a large portion of fraud losses if they do not comply with EMV specifications. However, the market response has been slow so far due to large investments needed in implementing the EMV compliant programs. As card business transactions increase, so too do frauds. Clearly, global networking presents as many new opportunities for criminals as it does for businesses. While offering numerous advantages and opening up new channels for transaction business, the internet has also brought in increased probability of fraud in credit card transactions. The good news is that technology for preventing credit card frauds is also improving many folds with passage of time. Reducing cost of computing is helping in introducing complex systems, which can analyze a fraudulent transaction in a matter of fraction of a second. It is equally important to identify the right segment of transactions, which should be subject to review, as every transaction does not have the same amount of risk associated with it. Finding the optimally balanced ‘total cost of fraud’ and other measures outlined in this article can assist acquiring and issuing banks in combating frauds more efficiently.

# CHAPTER THREE: METHODOLOGY

## 3.1 INTRODUCTION

This chapter comprises of the data collection methods aimed to be used in carrying out the process of project development and also the software development life cycle, this includes the software models and other documentation. Designing of the systems aims to focus on the users’ need and ensure the system is user friendly and interactive. I propose to use the waterfall model during the software development phase. The project progressively gains more complexity and a broader feature set until the final system is complete.

## 3.2 Data Collection

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.

### 3.3 Data Analysis

Analysis can be done using many ways but it is limited to the amount of data set size. It is unable to visualize large amount of data at a time. This project credit card fraud detection analysis overcomes the above mentioned issue in an efficient way by using R programming language. This proposed system can analyze and visualize large amount of data. In this project firstly the credit card data from the banks databases are collected. The credit card data which is collected is then converted into a pictorial format that may be either pie chart or bar chart. It is then reported as an analysis report and visualization report. However, the dataset from Kaggle is highly unbalanced with most transactions being legit and only a few being fraudulent. In this case, we used balancing techniques to balance the data set for efficient learning. The dataset was divided into a trained dataset and a test dataset. The techniques used were:

* Random Over-Sampling (ROS): here, the fraudulent transactions are duplicated to increase fraudulent cases. This method wasn’t the ideal balancing technique since it created duplicates of fraudulent cases and caused overlying of points in the plot.
* Random Under-Sampling (RUS): here, the legit cases are cut down to ensure balance between legit and fraudulent cases. Similarly, this method wasn’t upheld due to the fact that important data was discarded.
* Synthetic Minority Over-Sampling Technique (SMOTE): this method was best suited to tackle the unbalanced data problem. The general idea of this method is to artificially generate new examples of the minority class using the nearest neighbours of these cases. Furthermore, the majority class examples are also under-sampled leading to a more balanced dataset.

### 3.3.1 Pie chart of the dataset

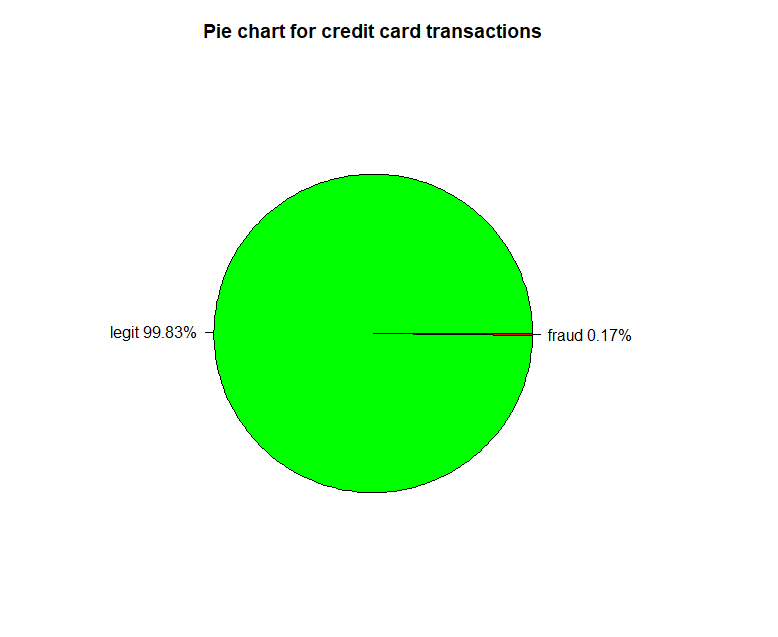


Figure 1: Pie chart for credit card transactions

### 3.3.2 Transactions distribution of original dataset

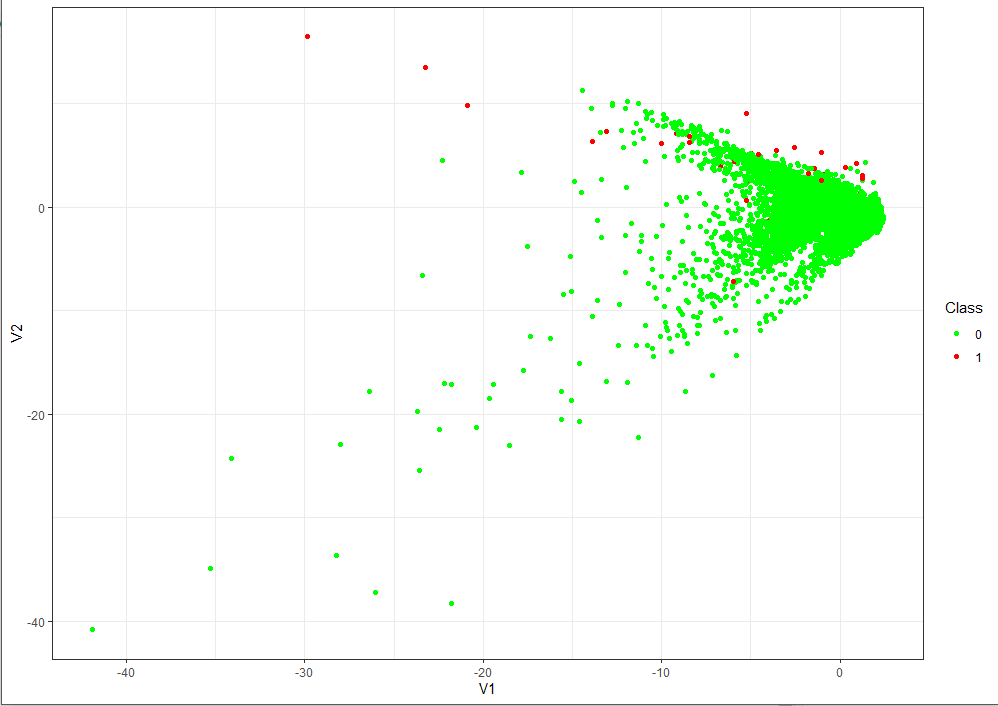


Figure 2: Scatter plot for credit card transactions dataset

### 3.3.3 Random over-sampling balancing technique output

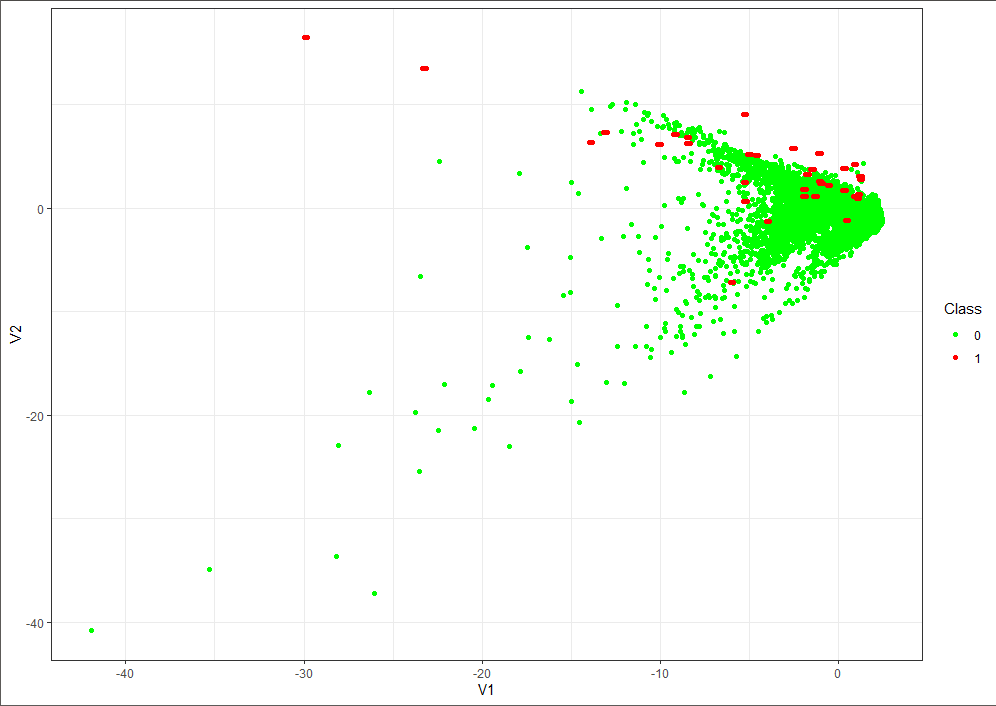


Figure 3: Scatter plot for ROS output

### 3.3.4 Random under-sampling balancing technique output

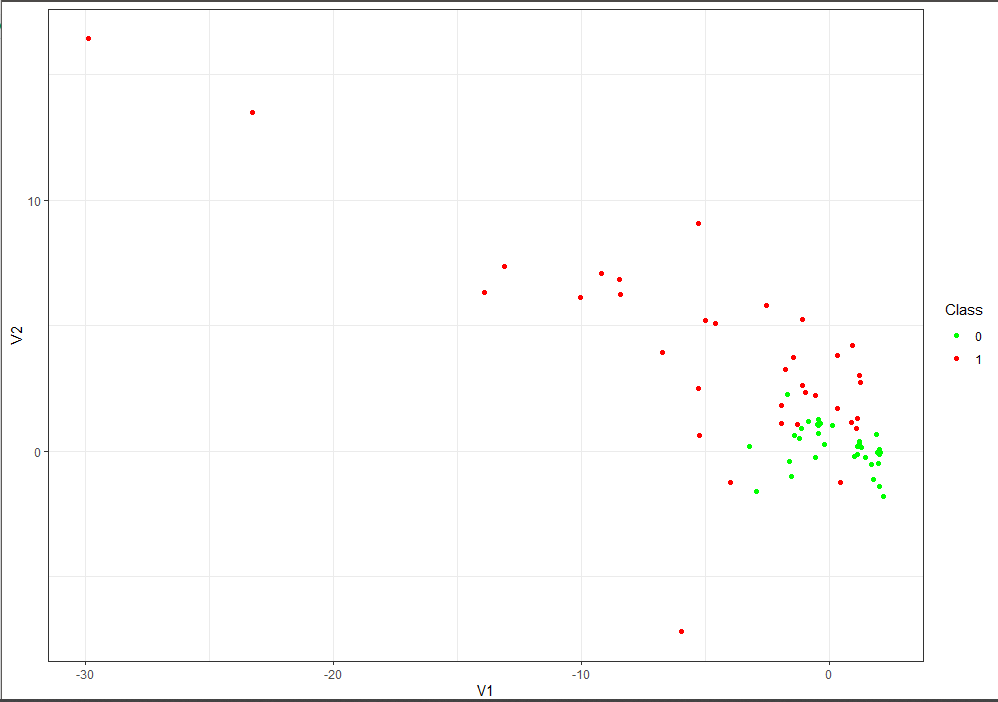


Figure 4: Scatter plot for RUS output

### 3.3.5 Synthetic Minority Over-sampling Technique (SMOTE) balancing output

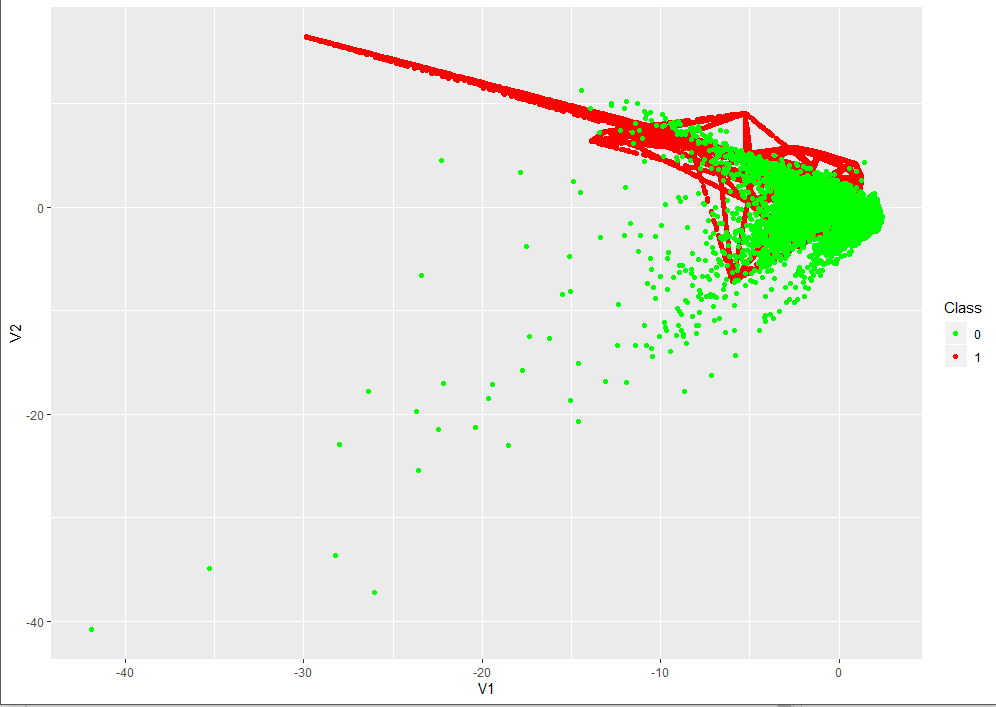


Figure 5: Scatter plot for SMOTE output

### 3.4 Tools

The project uses R programming language in RStudio IDE. The dataset is also implemented by Microsoft Excel.

# CHAPTER FOUR: SYSTEM MODELLING

## 4.1 Introduction

This chapter comprises of functional and nonfunctional requirements (requirement analysis) as well as the representations of how the system works.

## 4.2 Requirements Analysis

These includes the functional and non-functional requirements.

### 4.2.1 Functional Requirements

The functional requirements define the capabilities and functions the system performed successfully

1. Balancing of the dataset and representing it visually.
2. Predicting fraudulent transactions from the dataset.

### Non Functional Requirements

The non-functional requirements define or describe attributes such as security, reliability, maintainability, scalability and usability that the system has.

1. Accuracy – the system should ensure accuracy in predicting fraudulent transactions and differentiate them from legit ones.
2. Efficiency – the system uses adequate resources and does not over-utilize computer resources.
3. Fault tolerance – the system is designed to withstand any faults either from the framework and/or from other modules.
4. Maintainability – the system can be evolved and maintained with ease to ensure it meets rising needs.
5. Reliability – the system is always ready for use at whatever time that its user might require to use it.

## 4.4 System Analysis

### 4.4.1 Use case diagram

This is the diagram that is used to describe a set of actions that some systems should or can perform in collaboration with one or more external users of the system (actors).

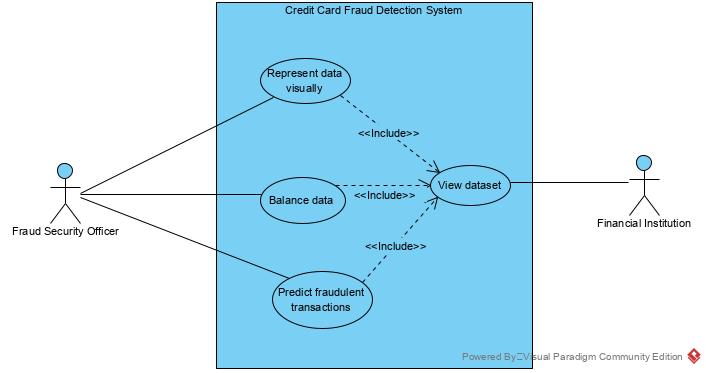


Figure 6: use case diagram

### 4.4.2 Sequence Diagram

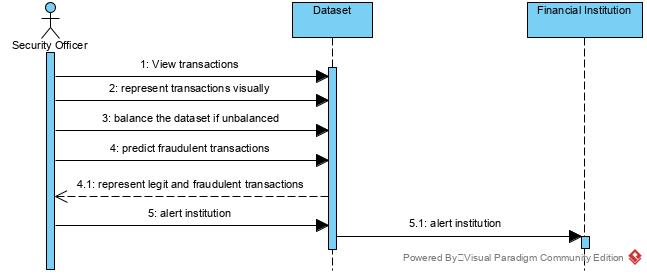


Figure 7: Sequence Diagram

### 4.4.3 Class diagram

This is a type of static structure diagram that describes the structure of a system by showing the system’s classes, their attributes, operations and the relationship among objects.

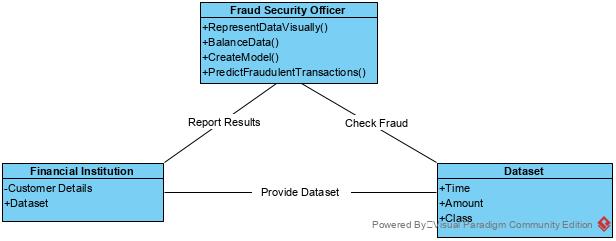


Figure 8: class diagram

### 4.4.4 Data flow diagram

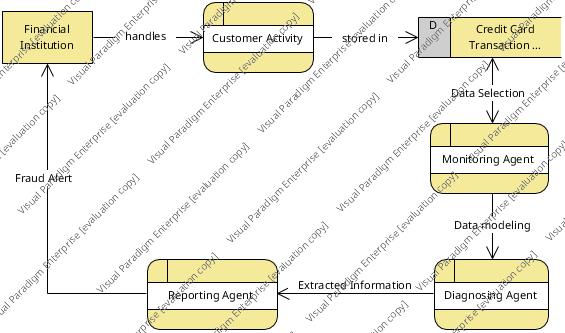


Figure 9: Data flow diagram

# CHAPTER FIVE: SYSTEM IMPLEMENTATION

## 5.0 Introduction

This chapter addresses mainly on the tools used for developing the new system.

## 5.1 Tools used for coding

In this stage, the tools I have used are: R language, RStudio, and Ms Excel

**R:** it is a programming language used mainly for statistical computing. Due to its extensive catalog, I used R for my project since it contains wonderful machine learning algorithms.

**RStudio:** this is the integrated development environment for R that I used to develop my system.

**Microsoft Excel:** this is where my dataset was held for the credit card transactions from Kaggle.

## 5.2 System testing and Implementation

### 5.2.1 Objectives

The major objectives of system testing and implementation are to check the developed system and implement. This section also provides the different strategies and implementation plan required for the purpose of testing and implementation.

### 5.2.2 Testing Overview

In software engineering, system testing plays an important role for the delivery of the project or the system. Every developed system is checked in order to find out the programming bugs. Meanwhile, testing cannot guarantee the total bug free system. Hence, the basic objectives of the system is to find out the bugs solve them for the bugs free system.

### 5.2.3 Testing strategy

For this system testing the strategy test plan contains a number of test cases followed by test log with test results.

### 5.2.4 Test plan

Test Plan Consists the different test cases prepare in order to test the system. The test plan with the test cases is mentioned below:

**Use case Test Plan**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test case | Use case | Objective |
|  |  | Import dataset | *To check if the transactions dataset can be successfully imported.* |
|  |  | Create train and test datasets | *To split the dataset into two parts; train and test subsets.* |
|  |  | Balance dataset | *To ensure the dataset is evenly balanced between legit and fraudulent transactions.* |
|  |  | Build the prediction model | *To build our predictive model for credit card fraud detection.* |
|  |  | Prediction | *To ensure the model created is efficient and working.* |

## 5.3 Implementation Plan

After completing the testing phase, software product is ready to be implemented in the application area. The implementation plan for the system is described below.

### 5.3.1 Process in Implementation plan

1. Identification of tools to implement the project.
2. Installation of the system.
3. User training.

**System Specification**

**Hardware Required**

* Pentium IV or similar processor
* 250 GB of Hard Disk space
* 4 GB RAM or Above

**Software Requirements**

* R 2.11.1 or higher
* RStudio IDE
* Required packages e.g. caret, caTools etc.
* Ms Excel

# CHAPTER SIX: LIMITATION, CONCLUSION AND RECOMMENDATION.

## 6.0 Introduction

This chapter concludes this project by highlighting the limitations of the project, conclusion and recommendations to be implemented in the future for the betterment of the system. It is majorly based on my experience while developing the system and my future changes to the system so as to fit the changing requirements of users.

## 6.1 Limitations from the project

During the time the system was being engineered, I faced challenges like difficulty in balancing the highly unevenly distributed dataset. Also, identifying the best model to use was rather time consuming.

## 6.2 Conclusion

Credit card fraud is without a doubt an act of criminal dishonesty. This proposal 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. The algorithm does reach over 99% accuracy and this high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of fraudulent transactions. 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. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

## 6.3 Recommendations

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.

# References

[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] “Credit Card Fraud Detection” by- Intellipaat; published by Intellipaat, India.

[3] “Article on Credit Card Fraud Detection” –by DataCamp published on 2019, [www.datacamp.com](http://www.datacamp.com)

[4] 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.

[5] “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

[6] “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 [