Application of Mahalanobis Distance in a Peer Profiling Model for Fraud Detection

By

Kimata, Larry Mutuku

152956

Supervisor: Dr. Allan Omondi

Master of Science in Information Technology

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya

2024

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a master's degree by this or any other university. This work contains no material previously published or written by another person except where the due reference is made in the work itself.

Student's Signature:

----------[Signature]

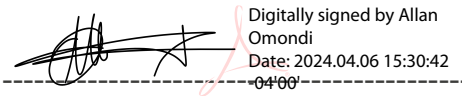
5TH APRIL 2024

----------[Date]

Approval

This work has been reviewed and approved (*for examination*) by:

Supervisor's Signature:



----------[Signature]

06-APR-2024

----------[Date]

Abstract

With the increasing volume of mobile and internet banking money transactions in today's era, fraud detection has become a major concern in ensuring a secure and trustworthy financial system. The rise in both digital money transactions has led to convenience and posed risks in opening gaps for fraudulent activities. Due to the changes in the digital era, fraudsters are looking for new techniques to exploit vulnerabilities in digital banking systems which leads to financial losses. The failure to stop fraud can lower the consumer benefit and financial inclusion gains in the businesses. There is a need to implement applications that accurately detect fraud as digital transactions in both mobile and internet banking has become the most used for fraud and other criminal activity.

The traditional methods involve rule-based systems to sort the data depending on the scenario that a human will term as a fraud. The human experts in the fraud domain will explore different transactions and customer accounts to note which ways swindlers use to perform suspicious activities. This will be then replicated as a rule in case that is done again the transaction will be termed as suspicious. There are a lot of limitations in this as there are infinite rules to be created to detect fraudulent activities. The systems that use have limitations that hinder efficient detection thus limited adaptability.

Current systems and approaches use different machine learning techniques that are based on pattern recognition which work well but the models must be retrained after a short time as they have a limited shelf life. The other problem is that swindlers and fraudsters also use machine learning in their arsenal of tricks to perform fraud; since they are smart people thus eloping fraud tool defenses as they understand how the pattern recognition models work thus fraud happening. This is because fraud evolves, the fraudsters evolve, and technology advances thus a better need for fraud detection ways.

The goal is to develop a tool that incorporates the use of the peer profiling method through clustering. The method involves the grouping of a dataset into different groups based on similar attributes and features of the peers. Afterward, new transactions can be classified from where the peer resides and use mahalanobis distance to calculate if the distance is away from the peer to be classified as an outlier. This proposed research aims to use the peer-based profiling and clustering algorithm to improve the accuracy of fraud detection. This will assist in reducing false positives and increase the accuracy of the frauds detected compared to the traditional way of detecting fraudulent activities.

Keywords: peer profiling, mahalanobis distance, fraud detection

Table of Contents

Declaration and Approval II
Abstract III
1 Chapter 1: Introduction 1
1.1 Background 1
1.2 Problem Statement 2
1.3 Research Objectives 2
1.3.1 General Objective 2
1.3.2 Specific Objectives 2
1.4 Research Questions 3
1.5 Hypothesis 3
1.6 Justification 3
1.7 Scope 3
2 Chapter2: Literature Review 4
2.1 Introduction 4
2.2 Theoretical Framework 4
2.2.1 Fraud Triangle Theory 4
2.3 Empirical Framework 6
2.4 Comparison of the related works to this research 7
2.5 Comparison with other Distance Measures 9
2.5.1 Euclidean Distance 9
2.5.2 Manhattan Distance 10
2.6 Software Development Framework 10
2.6.1 Agile Software Framework 10
2.6.2 Bootstrap Framework(Front-end Framework) 11
2.6.3 Scikit-learn Framework 11
2.7 Software Architecture 11
2.7.1 Model View Controller Architecture (MVC) 11
2.8 Review of the algorithms to be used 13
2.8.1 Peer Group Analysis 13
2.8.2 Mahalanobis distance 13
2.9 Research Gaps 15
2.10 Conceptual Framework 16
3 Chapter3: Research Methodology 18

3.1 Overview 18
3.2 Research Design 18
3.2.1 Type of Research 18
3.2.2 Type of Research Data 18
3.2.3 Test Data 19
3.2.4 Test Bed 20
3.2.5 Experiment Procedure 20
3.3 Systems Development 21
3.3.1 Agile Software Development Methodology 21
3.3.2 Data Collection 22
3.3.3 System Requirements 22
3.3.4 System Design 23
3.3.5 System Development Phase 23
3.3.6 System Evaluation and Testing 24
3.4 Research Quality 24
3.4.1 Reliability 24
3.4.2 Validity 25
3.5 Ethical Considerations 25
3.6 Conclusion 26
4 Chapter4: System Analysis and Design 27
4.0 Introduction 27
4.1 Dataset Description 28
4.1.1 Count of Transactions 28
4.1.2 Fraudulent and Genuine transactions counts 29
4.2 Requirement Analysis 30
4.2.1 Functional Requirements 30
4.2.2 Non-Functional Requirements 30
4.3 System Architecture 31
4.4 System Design 33
4.4.1 Use Case Diagrams and Activity Diagrams 33
4.4.2 Sequence Diagram 38
4.4.3 Database Schema 41
5 Chapter 5: System Implementation and Testing 44
5.1 Introduction 44

5.2 System Implementation 44
5.2.1 Mahalanobis Distance Implementation 44
5.2.2 Comparison with other Distance Measures 46
5.2.3 Summary of the comparison with other distance measures 52
5.2.4 Model Implementation and Testing 52
5.2.5 Web Management App Implementation 60
5.2.6 Banking App Implementation 63
5.2.7 Business Analyst Implementation 67
5.3 System Testing 70
5.3.1 Functional Testing 70
5.3.2 Validation Testing 71
5.3.3 Usability Testing 72
5.3.4 Compatibility Testing 72
6 Chapter6: Discussions and Key findings 73
6.1 Introduction 73
6.2 Primary Discoveries and Achievements 73
6.3 Exploration of the Research Objectives 74
6.4 Evaluation of the Prototype 76
7 Chapter7: Recommendations, Future Work and Conclusions 78
7.1 Recommendations 78
7.2 Future Work 78
7.3 Conclusions 79
References 80
Appendix A: Dataset Sample Portion 82
Appendix B: Turnitin Originality Report 83
Appendix C: Consent Form for Participation in the Research Study 84
Appendix D: Interview Form used to achieve Specific Objective one 86
Appendix E: SU-ISERC Ethical Approval 87
Appendix F: NACOSTI Research License 88

List of Figures

Figure 1: Fraud Triangle Theory....4
Figure 2: Fraud Diamond Theory....5
Figure 3 Euclidean Distance 1 Picture....9
Figure 4 Euclidean Distance 2 Picture....9
Figure 5 Manhattan Distance Picture....10
Figure 6: MVC Architecture....12
Figure 7: Conceptual Framework....16
Figure 8Agile System Development....21
Figure 9 Iterative Methodology for Agile....22
Figure 10 Transaction Distribution by Type....29
Figure 11 The Overall Architecture of the Prototype....32
Figure 12 Use case Diagram Prototype for the whole Prototype....33
Figure 13 Activity Diagram for Account Creation....34
Figure 14 Activity Diagram for customer performing transactions....35
Figure 15 Activity Diagram for business analyst reviewing cases....36
Figure 16 Activity Diagram for Business Analyst marks cases as Fraud or Genuine....37
Figure 17 Sequence Diagram for Admin Interaction with Prototype....38
Figure 18 Sequence Diagram for Customer Interaction with Prototype....39
Figure 19 Sequence Diagram for Business Analyst Interaction with the prototype....40
Figure 20 Main Database Schema....41
Figure 21 Transactions Table Schema....41
Figure 22 Transactions Table View Records....41
Figure 23 The Business Analysts Table schema....42
Figure 24 The Business Analysts Records View....42
Figure 25 The Customers Table Schema....43
Figure 26 The customers Table Records View....43

List of Tables

Table 1: Tabular Summary of Related Works 8
Table 2 Summary of Variables 20
Table 3 Summary of the Variables used 27
Table 4 Summary of the Comparison with Other Distance Measures 52
Table 5 Functional Testing: Banking App User Transacts 70
Table 6 Functional Testing: Web Application User Access 71
Table 7 Validation Testing 71
Table 8 Usability Testing 72
Table 9 Compatibility Testing 72

List of Equations

Equation 1: MD Equation 1 14
Equation 2: MD Equation 2 14
Equation 3: MD Equation 3 14
Equation 4: MD Equation 4 14
Equation 5: MD Equation 5 14

Abbreviations and Acronyms

ACFE – Association of Certified Fraud Examiners

API – Application Programming Interface

APP - Application

BA – Business Analyst

FTT – Fraud Triangle Theory

FDT – Fraud Diamond Theory

IMF – International Monetary Fund

MD – Mahalanobis Distance

NACOSTI – National Commission for Science, Technology, and Innovation

PGA – Peer Group Analysis

REST – Representational State Transfer

SU-ISERC – Strathmore University Institutional Scientific and Ethical Review Committee

1 Chapter 1: Introduction

1.1 Background

Digital transactions in both internet banking and mobile money have become a consequential challenge for economies, businesses, and economies globally which has imposed to negative economic costs due to fraudulent activities that encompass those channels. This has also led to undermining of trust in financial services industry due to monetary loses thus hampering sustainable development. According to the International Monetary Fund (IMF), illicit financial flows can have a significant impact on the economic stability and the global financial system such as lower tax receipts, distort competition, inflate prices, and reduce government revenue (IMF,2023). These illicit financial flows can be money laundering, forms of fraud, embezzlement, and tax evasion.

The mitigation of fraud risks is important to organizations and financial institutions to safeguard their assets. Authorities and financial institutions including money transfer companies are continuously looking for new ways to identify and stop these frauds just as fraudsters are coming up with new strategies to stay a step ahead. Most governments have set up fraud taskforces to combat this, but it has not been enough due to the evolving nature of swindlers.

Technological advancements which lead in introduction of smartphones has led a spike in usage of digital banking transactions. This is due to the convenience, speed and accessibility compared to the old way of financial institutions where transactions must be made physically at the bank. This has also fuelled a huge growth for both the economy and many businesses. This has also an impact on security as there is a need of cultivating trust by building dependable and trustworthy digital services which are the enablers of mobile money transactions (Choudhury, A. et al, 2022).

Due to the increase of digital banking transactions globally; money launderers, attackers and fraudsters all view it as the desirable target. Although, it is compulsory for service providers and financial institutions in many nations to report transaction fraud and money laundering, the relatively new mobile and internet banking money markets are not yet fully considered many of the countries' laws and regulations. The risk is that this increases with lack of oversight and an effective fraud tool which can detect this. The use of this digital channels has significantly provided unbanked populations to acquiring access to financial instruments and different money channels. By using these services, for an example is that people can have access to informational services such as transactional services, balance inquiries in bank accounts and purchase of goods and services.

According to Wronka (2022), Despite of some detection services now being out in place, investigating and monitoring money transactions to pinpoint the questionable activities remains challenging. This is because fraudsters are using the digital channels which are less

regulated by traditional financial institutions to perform the frauds, this calls into sustainability, trust, and integrity question of mobile money transactions.

1.2 Problem Statement

The Association of Certified Fraud Examiners (ACFE) estimates that organizations lose close to 5% of their annual revenues to fraud. These costs impact the profitability and affect the financial stability of businesses while affecting the stakeholder's trust in the organization operations (ACFE,2016).

Mobile money and internet banking transactions have become widely used as the way of doing financial transactions. The security and integrity of digital money systems are however, seriously threatened by fraudulent activities that are attracted due to their popularity thus leading to financial loses (Buku et al.,2017).

Current efforts done by the financial institutions is the use of rule based and pattern recognition techniques. Rule based systems have been the traditional approach which right now has limited adaptability due to traversing several large datasets with extremely complex rules. A study by Burmeister et al. (2014) states that the rule-based systems are far from perfect in their performance due to the limited adaptability. The use of pattern-based recognition techniques which are the current norm is a good machine learning fraud detection but there is a need to retrain the models after a very short time which is a limitation as it's not evolutionary. With the digital era, fraudsters and swindlers also understand the logic which machine learning tools use thus they are smart and able to use their tricks to elope the pattern-based models. According to Kaur et al. (2019) fraudsters may actively study the rules and patterns used by fraud detection tools to identify weaknesses and exploit them.

The proposed research will have the application of peer profiling technique and machine learning can assist to better improve fraud detection with better precision and accuracy due to the evolutionary model that would be built.

1.3 Research Objectives

1.3.1 General Objective

To design and develop an application that can predict fraud from the digital banking transactions data using peer profiling and mahalanobis multivariate analysis distance.

1.3.2 Specific Objectives

i. To investigate the challenges faced by financial institutions due to fraudulent activities.
ii. To review the current used methods in the internet banking fraud detection.
iii. To train the model by developing peer profiling and use of Mahalanobis distance multivariate analysis in the anomaly detection.
iv. To validate the model using mobile money transactions' datasets and comparing the fraud label on the data.

1.4 Research Questions

i. What are the existing techniques used in fraud detection in financial systems?
ii. How effective are the current methods used in digital banking fraud detection?
iii. How can one design and develop a model for fraud detection by using per profiling and machine learning?
iv. How effective is the proposed model in fraud detection of digital banking transactions?

1.5 Hypothesis

The null hypothesis is that there is no significant difference in the effectiveness of detection of fraud using peer profiling and use of mahalanobis distance in securing digital banking transactions compared to the current methods.

1.6 Justification

Financial transactions in this era are increasingly shifting towards digital services and platforms but there is a need for an assurance that customers' funds are protected against fraudulent activities inorder to safeguard their assets.

Due to the increased popularity and demand for the digital banking transactions, there is a significant need for an evolutionary and accurate fraud detection tool. The existence of such a system would ensure security and integrity of money systems threatened by fraudulent activities. This will also contribute to the growth and stability of the digital banking ecosystems.

1.7 Scope

This research will be focused on identifying fraudulent transactions within data for financial institutions. The study will look at the use of peer profiling and machine learning to achieve the fraud detection of the mobile money transactions.

2 Chapter2: Literature Review

2.1 Introduction

This chapter will give an overview of the concept and investigation of the research problem comprehensively; an empirical framework is presented to show evidence-based research revolving the fraud detection using a peer profiling model which uses Mahalanobis multivariate distance analysis. Significant publications done by researchers are further reviewed to show the theoretical framework.

The challenges faced by financial institutions due to fraud are losses and financial impact, reputation damage, severe penalties and legal actions from regulatory compliance and data security concerns.

2.2 Theoretical Framework

2.2.1 Fraud Triangle Theory

This section describes the fraud triangle theory (FTT) which is an important theory that will assist to show the conditions and components that lead to people to commit fraud. The study of fraud can be understood using this theory which was proposed by Donald R. Cressey. It mentions that there are factors and elements that usher to people in committing fraud which is an unethical activity. The elements include pressure, opportunity, and rationalization (Sánchez et al., 2021).

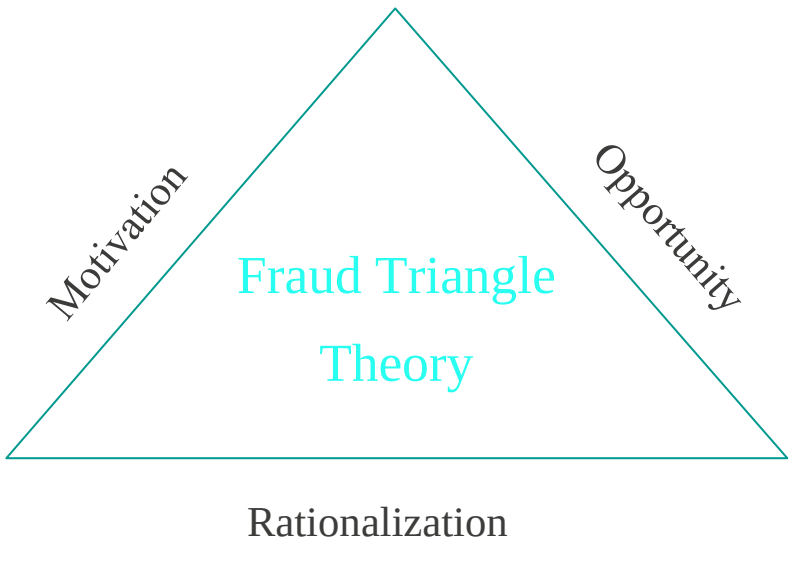


Figure 1: Fraud Triangle Theory

The first element, which is pressure, is related to the motivation and greed for fraudulent actions of a person. Financial stress is a common factor that plays a big role in provoking the desire for greed. Perceived pressure may occur will all employees at any level in a company as this can occur due to various reasons, specifically economic pressure.

The second element is opportunity; fraud majorly occurs in financial systems where there is a lack of strong internal controls and governance to inhibit this. Fraudsters will proceed and exploit this as an opportunity has been created to commit the swindling activities to gain benefits. Financial institutions which lack appropriate and robust internal controls with governance are culprits to financial risk and losses as an opportunity and loophole are there for the fraudsters.

The third element, referred to as rationalization, occurs when an individual justifies their fraudulent activities to themselves as okay; since one convinces themselves what they are doing is not wrong or they deserve the benefit they will gain from the dishonesty act, thus making their illegal actions seem acceptable and accounted of.

Despite the alignment of the three vertices of the fraud triangle theory, individuals are unlikely to perpetrate fraud unless they have the capacity to do so. The fourth element of capability was introduced from the extension of FTT and added to have a fraud diamond theory(FDT). This can be considered as the fourth vertex, as the potential fraudster must have the skillset and ability to commit fraud. Ruankaew (2016) states that this may be due to a person's position or function within a company that may give him or her the ability to create or exploit the opportunity for fraud; intelligence to exploit the internal controls system is key to note in this fourth vertex and ability to understand how the system works as they are able to identify the weaknesses and avoid getting caught.

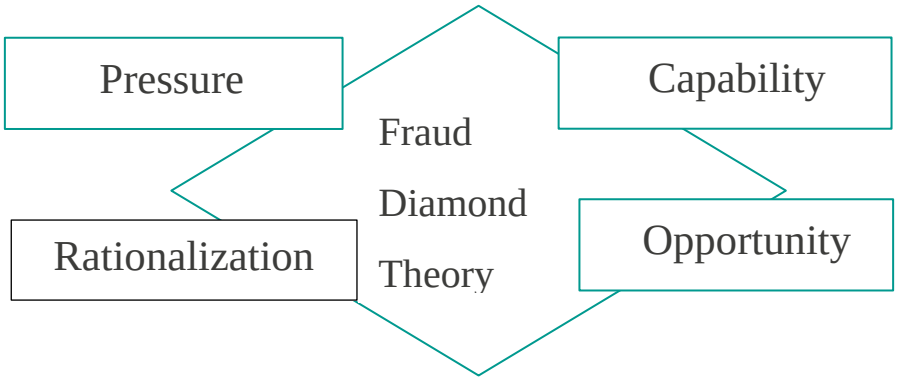


Figure 2: Fraud Diamond Theory

Fraud detection is complex as it requires a sophisticated model which would require interpretation of human behaviors. The FTT and FDT theory gives us an overview of the theoretical framework and the vertex that motivate fraudsters. In conclusion, there is a need to develop a fraud tool that will have strong, robust controls in financial institutions to fill the gap in the loopholes and address the components addressed in this theoretical framework. The application of Mahalanobis distance in a peer profiling model will assist in filling the gap

of the FTT theory through increased enforcement, improved auditing standards, and a sophisticated fraud detection tool.

2.3 Empirical Framework

Plastic card fraud detection using peer group analysis (David et al., 2008)

All the authors were David J. Weston, David J. Hand, Niall M. Adams, Christopher Whitrow, Piotr Juszczak

This evidence-based research used peer group analysis which is a technique that involves developing peers and reviewing transactions that strongly deviate from the peer group and are flagged as potentially fraudulent. This research used peer group analysis and Euclidean distance as its main techniques.

In this research, there were some weaknesses, such as no hyperparameter tuning for the longevity of the model, and the low accuracy outcome, involved using Euclidean distance which is not the best for multivariate analysis. Mahalanobis distance is preferred to Euclidean distance because it considers point distributions, that is, correlations. The research achieved an accuracy of 79.34%.

Implementing peer group analysis within a track and trace system to detect potential fraud (Foo et al., 2014)

All the authors were Foo Chi Hui, Venkaiah Chowdary Koneru, Norazman Mat Ali, Safurah Harun

In this research, the authors worked on reviewing a method that involves the use of PGA (Peer Group Analysis) to detect anomalies in event transactions in a supply chain management and analysis to be able to understand where disputes, frauds, and thefts happen.

For future work, the authors advised that there is a lot more to refine within peer group analysis, and a plan to develop a dynamic peer group building would be beneficial to future works. Furthermore, the research notes that further works reviewing the use of peer group analysis for real-world problems will be important too.

As discussed above, the drawbacks of this research were low accuracy achieved, and no dynamic peer group building employed.

Multisource fusion for anomaly detection using across domain and across time peer group consistency checks (Hoda et al., 2014)

The authors involved in this research were Hoda Eldardiry, Kumar Sricharan, Juan Liu, John Hanley, Bob Price, Oliver Brdiczka, and Eugene Bart

In this study, the authors worked on robust anomaly detection that detects anomalies in heterogeneous data. The entities involved were people from observation data activities from multiple contexts of a company system. This was achieved by peer group analysis, mainly peer group clusters where different people departments such as system admins, finance, and engineers were grouped, and their data checked for inconsistencies within a distribution to spot the anomalies.

This research used the Markov model, peer group analysis, and automatic entropy-based threshold selection. The only drawbacks of this research were a medium accuracy of 79%, and peers were only built on clusters level.

Peer Group Analysis – Local Anomaly Detection in Longitudinal Data (Richard et al., 2001)

The authors who worked on this research were, Richard J. Bolton and David J. Hand.

This works explains how peer group analysis can be used in anomaly detection for longitudinal data. This will be an important piece of research for my work as it outlines some of the key concepts that I will use.

In this research, the techniques used are peer group analysis and Poisson distribution. The drawbacks of this are the accuracy achieved which was less than 70%, a cross-validation issue since there was an overfitting within the models created, and a peer group building issue.

2.4 Comparison of the related works to this research

The different four related works have embraced the use of peer profiling technique in there researches. Although, the works are not directed to the mobile money and internet Banking transactions, they have all achieved the fraud detection of the certain scenarios where the research revolves around.

This research will borrow on the peer profiling technique concept and use it in mobile money and internet Banking transactions. The works cited showed how peer profiling could be used to detect fraudulent activities on the concepts reviewed regardless of the scenarios.

Below is a Summary of the four related works in a tabular format : -

	Authors	Title	Technique Used	Results	Weaknesses/Comments
1	David J. Weston, David J. Hand, Niall M. Adams, Christopher Whitrow, Piotr Juszczak	Plastic card fraud detection using peer group analysis (David et al., 2008)	+Peer group analysis +Euclidean Distance	Accuracy:79.34%	+Peer groups building issue +The authors are of the opinion that the Euclidean distance is not scale-invariant, distances that are computed might be skewed leading to wrong calculations thus it did not work well. +Medium Accuracy
2	Foo Chi Hui, Venkaiah Chowdary Koneru, Norazman Mat Ali, Safurah Harun	Implementing peer group analysis within a track and trace system to detect potential frauds (Foo et al., 2014)	+Peer group analysis	Low accuracy	+Low accuracy +No dynamic peer group building employed
3	Hoda Eldardiry, Kumar Sricharan, Juan Liu, John Hanley, Bob Price, Oliver Brdiczka, and Eugene Bart	Multisource fusion for anomaly detection using across domain and across time peer group consistency checks (Hoda et al., 2014)	+Markov Model +Peer group analysis +Automatic entropy-based threshold selection	Accuracy:79%	+Medium accuracy +Peer group building issue.
4	Richard J. Bolton and David J. Hand	Peer Group Analysis – Local Anomaly Detection in Longitudinal Data (Richard et al., 2001)	+Peer group Analysis +Poisson Distribution	Accuracy:<70%	+Low accuracy +Peer group building was an issue

Table 1: Tabular Summary of Related Works

2.5 Comparison with other Distance Measures

In this research, we will compare the results of Mahalanobis distance with other two distances to bring out the uniqueness of the MD and how it performs better. In this work we will compare the mahalanobis distance with Euclidean distance and Manhattan Distance. The distances have been briefly reviewed below which will be used in comparison to the MD.

2.5.1 Euclidean Distance

This distance is calculated by getting the distance between two points of the line segment between them in a multidimensional space(Liberti et. al., 2017).

Euclidean distance=$(x_2-x_1)_2+(y_2-y_1)_2$

Simply Euclidean distance can be calculated using the above equation and represented as below. The below show how we achieve the calculation by use of the Pythagoras theorem.

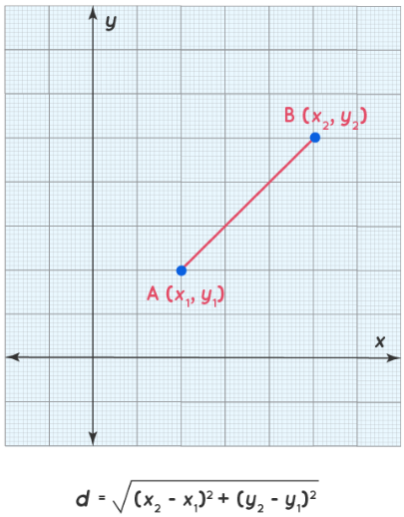


Figure 3 Euclidean Distance 1 Picture

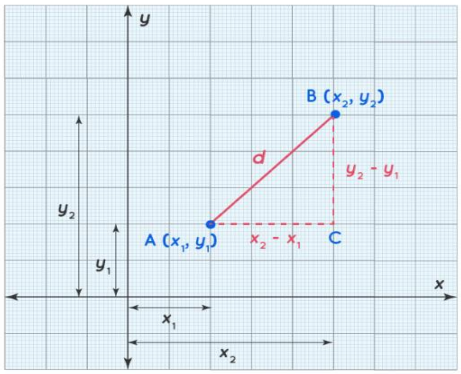


Figure 4 Euclidean Distance 2 Picture

AB2 = AC2 + BC2

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

$$d = \sqrt{[(x_2 2 - x_1 1)^2 + (y_2 2 - y_1 1)^2]}$$

2.5.2 Manhattan Distance

This is a distance metric that is used to measure distance between two points which considers only orthogonal(right-angle) movements to reach from one point to another.

In a 2D space, the distance can be calculated by getting the sum of the absolute differences of their coordinates.

Manhattan distance=$|x2-x1|+|y2-y1|$

Manhattan distance=$\sum i=1n|x2i-x1i|$

The distance can be represented as below:-

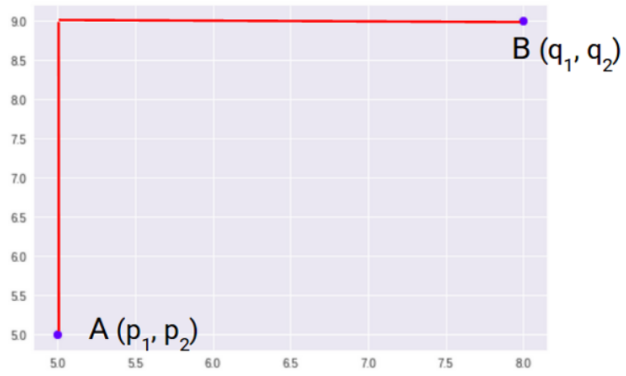


Figure 5 Manhattan Distance Picture

2.6 Software Development Framework

This section presents the software development framework in form of reviewing the technical details involved, the application of Mahalanobis distance in a peer profiling model for fraud detection will involve the use of the following:

2.6.1 Agile Software Framework

In this phase, the research will be implemented using the agile framework due to its iterative nature to customize elements to meet the unique set requirements checklist for the research. These elements of continuous planning, testing, integration, and continuous development will be incorporated as they are rules and practices tied to this framework. Mainly, I will use the rapid application development framework within the agile approach.

2.6.2 Bootstrap Framework(Front-end Framework)

This framework is primarily used in building user interfaces in web applications, dashboards and interactive components which makes it a unique framework for this research.

Unique features that are available and will be used are real-time updates, component reusability and state management.

It works by using the below architecture and view:-

- Component-based architecture: This project will be unique as by using this it will be decomposed into reusable and encapsulated component units.
- Virtual DOM: DOM refers to Document Object Model, this approach minimizes browser manipulation and enhancing performance.

2.6.3 Scikit-learn Framework

This is referred to as sklearn, which is a popular machine learning library for python. It is widely used in tasks such as data preprocessing, model selection, model training and model evaluation(Hao et. Al, 2019).

This is a valuable tool that will assist in data analysis, predictive modelling, and statistical analysis.

2.7 Software Architecture

This section briefly reviews the technical details for the overall application structure and behavior.

2.7.1 Model View Controller Architecture (MVC)

This is an architectural pattern that separates an application into three distinct logical components to handle specific development views of the application. This would assist in showing the interface between the model and the view.

For this, the MVC will assist in demonstrating the below in the specific logical components:-

Model: This will represent the data structure and logic for the peer group analysis, application of Mahalanobis distance calculation, and anomaly detection.

The peer group analysis component will be responsible for analyzing the data and grouping the specific users into peer groups based on similar behaviors and attributes. This phase will involve other data science techniques, such as data preprocessing, feature engineering, and clustering algorithms, to create distinct peer groups.

For the Mahalanobis distance calculation, it will also be within this component, where the distance will be calculated for each incoming transaction respective to its peer group. Anomaly detection will be enhanced by this component as it will detect fraudulent transactions that exceed a certain threshold based on the distance, thus flagged as potential fraud cases.

View: This will depict the frontend user interface where application users are able to log in and see the raised fraudulent transactions/hits. This will display the transaction data and presents the results of the fraud cases with an intuitive and user-friendly interface.

Controller: This will assist in handling user interactions and requests while managing the flow of data within the fraud detection model.

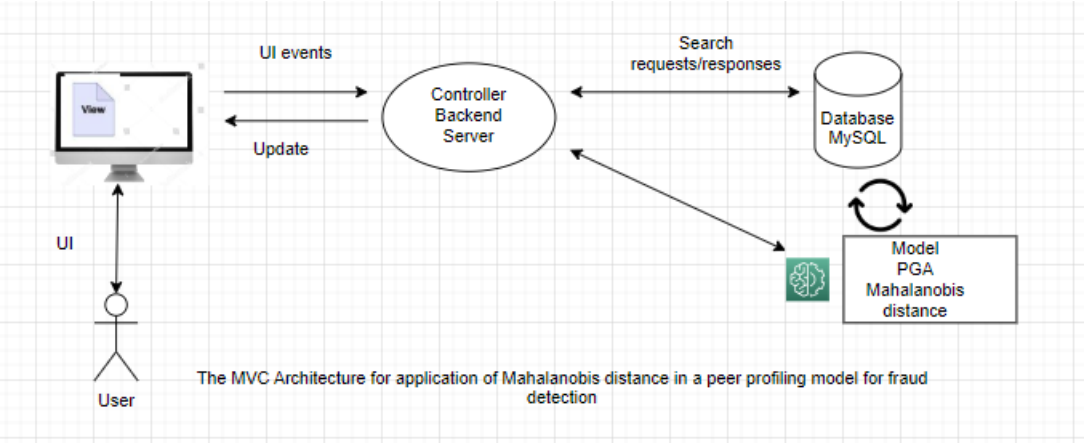


Figure 6: MVC Architecture

Below are the interactions for this MVC components:

The user will log in to the frontend(UI), which is the view, and interact with it to be able to view the data and the hits. The user request is sent to the controller, which in this case is the controller.

The backend server processes the received request and interacts with the database and model.

If this is based on a new transaction, the model performs the necessary calculation and returns the results to the backend server.

The backend server finally passes the results to the view, which displays transaction data and anomaly results to the user.

The MVC architecture assists in showing a distinct view of the architecture as it promotes modularity and flexibility in this fraud detection model.

2.8 Review of the algorithms to be used

2.8.1 Peer Group Analysis

The PGA technique involves the process of comparing entities that share similar characteristics. Before comparison, the data must be segmented into unique groups where the entities or users are like one another. This technique is used to monitor behavior over time in data mining applications.

The primary role of the PGA is to distinguish the anticipated pattern of behavior around the target chronology of events in terms f the behavior of the same objects and then discover any dissimilarity in evolution between the expected pattern and the target.

PGA analyses transaction data and groups users that behave in a distinct way depending on transaction history, and user profile attributes that match the peer group. Each user is compared with all other users in the data set, using different comparison criteria; patterns are identified, which are used to match the peer groups. The behavior of the created peer group is then summarized at each subsequent time point, and the behavior of the incoming test data transaction is compared with the summary of its peer group (Richard et al., 2001).

If, in a dataset, we have a specific number of M objects, and we must decide how many objects peer it contains so that we can create an algorithm to be used to create peer groups. The parameter npeer constructively manages the susceptibility of the peer group analysis. The most computationally thorough aspect of PGA is to find distinct attributes or features to determine the peer groups, as this is what leads to a highly accurate and efficient fraud detection model. In this research, it will involve using this peer group analysis concept and creating an algorithm that will be used in the peer profiles for fraud detection.

2.8.2 Mahalanobis distance

Mahalanobis distance (MD) refers to the measure of distance between a particular point P and a distribution. The technique was identified by P.C. Mahalanobis, the main definition of MD is that it was prompted by the unique problem of identifying similarities of skulls based on measurements. The unique idea of the MD is it brings out the aspect of how many standard deviations away P is from the mean D.

MD is computed as the square root of the difference vector's product, the inverse covariance matrix, and the difference vector's transpose. MD considers the scale, correlation, and shape of the variables and their linearly transformable. It is one of the best multivariate analysis techniques for finding outliers since it measures how far a point deviates from the mean along each primary component.

Mahalanobis distance is preferred over Euclidean distance(ED) because ED does not consider the scale or correlation of the variables, so it can be skewed by outliers or skewed distributions.

Etherington (2021) states that MD are based on the location and scatter of a multivariate distribution and can be used to calculate how far any point in space is from the center of this type of distribution. The variance-covariance and sample mean, minimum covariance determinant, and minimum volume ellipsoid are commonly used methods for computing multivariate location and scatter.

Wikipedia provides a summary of the Mahalanobis distance using the below definition which considers use of equations and formulas; Given a probability distribution Q on R^N , with mean u.

Equation 1: MD Equation 1

$$\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^\mathsf{T}$$

And positive-definite covariance matrix S, the MD of a point x ;

Equation 2: MD Equation 2

$$\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^\mathsf{T}$$

From Q is

Equation 3: MD Equation 3

$$d_M(\vec{x}, Q) = \sqrt{(\vec{x} - \vec{\mu})^\mathsf{T} S^{-1} (\vec{x} - \vec{\mu})}.$$

Given two points x and y in R^N, the MD between the in respect to Q is;

Equation 4: MD Equation 4

$$d_M(\vec{x}, \vec{y}; Q) = \sqrt{(\vec{x} - \vec{y})^\mathsf{T} S^{-1} (\vec{x} - \vec{y})}.$$

Which refers to that

Equation 5: MD Equation 5

$$d_M(\vec{x}, Q) = d_M(\vec{x}, \vec{\mu}; Q)$$

2.9 Research Gaps

The research intends to contribute to the advancement of fraud detection by fostering trust in digital financial services and enhancing the security of the established controls. This research will be able to bridge the following gaps:

I. Adoption of MD for anomaly/outlier detection: Since Mahalanobis distance is an effective metric for detecting outliers and anomalies detection, its application has not been extensively studied. This research seeks to validate the suitability of the Mahalanobis distance application in detecting fraudulent activities.

II. Utilizing peer group analysis for fraud detection: This technique is a unique approach to group different users with similar attributes, features, and behaviors, thus enabling a more accurate anomaly detection method. However, PGA application in view of fraud detection is relatively unexplored.

III. Handling Evolving fraudulent tactics and activities: Due to the increasing fraud, swindlers have found new ways of using machine learning to counter this and perform more fraudulent activities. This research is aimed at exploring methods to identify and adapt to evolving fraud patterns, thus ensuring that the model has a longevity feature and attribute.

IV. Handling skewed distribution: By using Mahalanobis distance in this research, it will be able to address the challenges of skewed distribution, which can lead to less accurate fraud models.

V. Optimal Peer Group Size and composition: By using the peer group analysis technique, we will be able to bridge a gap in identifying the most relevant transaction features, user attributes, peer group definition, and building and having the optimal size for peer groups.

2.10 Conceptual Framework

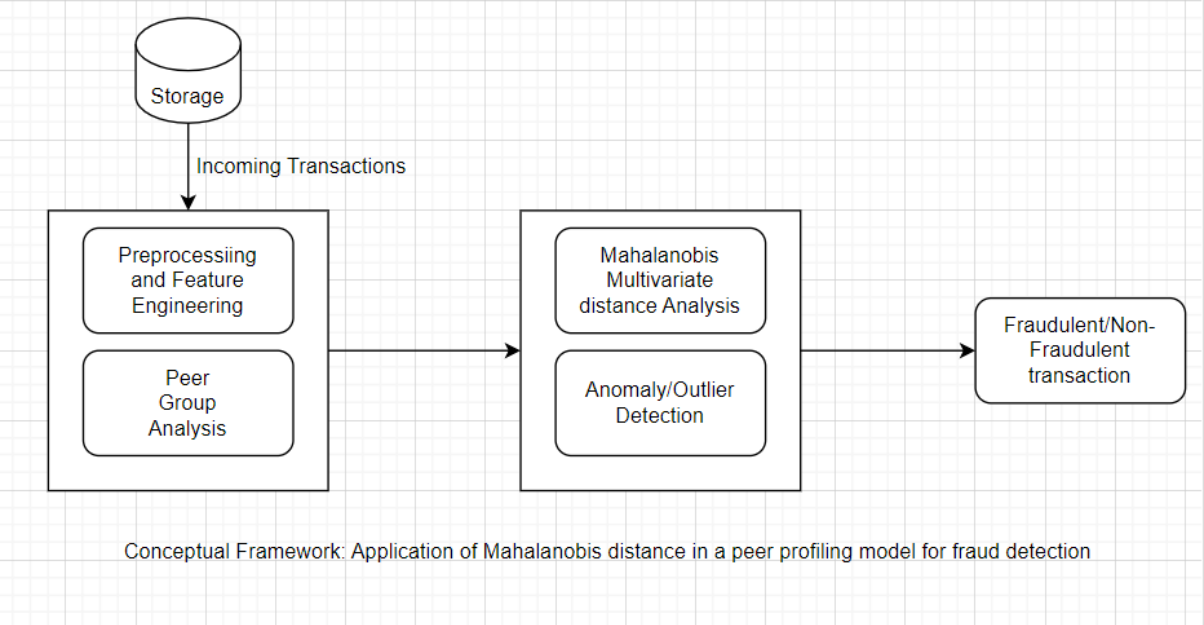


Figure 7: Conceptual Framework

Storage: This is a component that will store the data for transactions, both historical and newly incoming transactions, to be reviewed by the model.

Preprocessing: This component will house the steps taken to clean, transform and organize the data so that it can be used for analysis. It is a crucial step as it will transform the data to be ready for further modeling. The preprocessing tasks involve data cleaning, data transformation, and feature selection.

Feature Engineering: This will be a crucial step in preparing the data ready for peer group analysis. This will involve transforming the data into underlying patterns and relationships that present the data by generating the features. This will assist in developing more accurate and interpretable models, thus improving the detection of fraudulent activities.

Peer group Analysis: This is a technique used to compare a specific entity with a group of its peers based on similar behaviors, attributes, and features. The steps involved in peer group analysis are:

I. Identifying the peer group: Determine the group of entities that will be used for comparison that leads to peer group building.
II. Data Collection
III. Normalization: This will involve calculating ratios or relative values to standardize the data across the entities.
IV. Comparative analysis: The normalized data will be used to compare the performance of entity with its peers. Mahalanobis distance will be employed to measure the

difference between the points and center of distribution to determine the outliers which will be considered as fraudulent activities.

Mahalanobis distance: Application of MD to detect outliers/anomalies after the peer group analysis.

3 Chapter3: Research Methodology

3.1 Overview

For this chapter, this is guided by the proposed objectives of the research outlined in chapter one, the nature of the problem to be explored and research designs relevant to this work. The below activities compose the research methodology which in a nutshell are the sampling technique, to propose the methods of determining the items to be observed in research, the criteria by outlining tools to be used and use of statistical technique methods to further explain how the information collected is to be analyzed. Additionally, it also includes approaches in the analysis and design process(Wells,2009). This research will use the structured system analysis and design approach(SSAD). SSAD is suitable when one is approaching complex system requirements that require much analysis and planning. In the context of this research, designing a model that will have unique transaction patterns while interacting different components of the systems and an application of Mahalanobis distance, this approach will ensure that these complexities are well systematically addressed and documented.

3.2 Research Design

This research proposes to develop a web-based tool that will accept transactional data as input, process this data using a peer profiling algorithm, and use the analysis in the stored prediction model to check how far the weight is by using Mahalanobis distance inorder to predict whether a transaction is fraudulent or non-fraudulent.

3.2.1 Type of Research

The research design to be used in this study is applied research design because it is crucial to solve the practical problems and generate solutions that are directly applicable to a real-world scenario like fraud detection. Applying this research design will assist to address the problem of fraud detection as explained in the research.

3.2.2 Type of Research Data

Using a quantitative design will be crucial in the study as it will provide the empirical data with the analysis needed to build and validate the predictive models. This research data technique will assist in creating and refining a fraud predictive model that can accurately identify fraudulent transactions.

The data used for this research will be secondary which will involve use of Paysim Simulator to get the data to be used for this research. The methodology used in requirements gathering will be document review approach to make sure all the captured requirements are met.

To achieve objective one of this proposal which is to investigate the challenges faced by financial institutions due to fraudulent activities, we will be using primary data collection by using interviews to achieve this.

The sample size and sampling method will be guided by the need to gather information from experienced professionals in the fintech space who have great knowledge on the fraud landscape to get this data to assist us answer this.

Due to the above, this research will use purposive sampling. This method will involve selecting participants who are directly involved with fraud detection or compliance within the financial institutions. Since the topic in question needs more experience, this method will assist to provide most relevant and informed responses.

The sample size will involve considering several factors including the diversity of the financial institutions and the variability of the challenges faced by this organizations. Since this is important to answer the first objective of the research, a sample size of 30 to 50 respondents will be used to allow capture responses from a diverse range of professionals.

To demonstrate clear inclusion and exclusion criteria, from the above used sample size, the research will further make sure the below are met:-

- Demographic characteristics: This will involve considering age, gender, and education level to get a correct data collection criterion from the sample size.
- Inclusion and exclusion criteria Inclusion of professionals who have more experience in the fraud landscape and less to get their input which assist answer the questions in a broad way.
- Ethical considerations: Ensuring that the research adheres to ethical considerations and do not discriminate against any group or individual.

3.2.3 Test Data

The dataset to be generated will involve features such transaction amount, sender's transaction counts, destinations transaction count, residual balances, the sender account id, the destination account id, and transaction type used.

These variables will assist to build the profiles to be used in detecting the fraudulent transactions. The dataset will be split into two, into a training dataset which be used for this and a second dataset which will be used to validate the working of the model.

Below are the sample variables and their description.

Table 2 Summary of Variables

#	Variable Feature Name	Example format for the Variable Name	Further Description
1	step	5	Each step is an hour of time in real world. The largest number for step is 744 (the 30th day)
2	type	PAYMENT (Categorical variable)	Transaction types (CASHOUT ,CASH-IN, DEBIT, TRANSFER AND PAYMENT)
3	amount	8424.74	Transaction amount in local currency
4	nameOrig	C1000001725	Customer who initiated the transaction
5	oldbalanceOrig	351422.72	The initial balance of sender before the transaction
6	newbalanceOrig	257557.59	The new balance of sender after the transaction
7	nameDest	M1974356374	Customer/Merchant who received the transaction
8	oldbalanceDest	526950.37	The initial balance of receiver before the transaction
9	newbalanceDest	771436.84	The new balance of receiver after the transaction
10	isFraud	1 (Categorical variable)	The status of a transaction (0 as genuine and 1 as fraudulent)

3.2.4 Test Bed

The environment to be used in building the model will be using python, Sklearn, SciPy, pandas, and NumPy libraries to the development of the model. This will describe the overview of how the model will be working. Model evaluation is an integral part of this testing where cross validation and hyperparameter tuning will be used.

To perform a successful training environment, computational resources must be sufficient to make sure that the models are working well, this is in terms of CPU, RAM, and Storage.

3.2.5 Experiment Procedure

The below steps show a brief 3 steps that will incorporate to this research being successful; more steps will be detailed once we proceed with data analysis and statistical tests.

I. Data preprocessing, cleaning, and feature extraction.
II. Build a peer profiling algorithm depending on the features described above.

III. Use of application of Mahalanobis distance to calculate the distance between the new transaction and the historical peer distance. Using the distance set as threshold, define a transaction as fraudulent or non-fraudulent.

3.3 Systems Development

3.3.1 Agile Software Development Methodology

The below diagram shows the basic steps involved in an agile software development lifecycle; it is an iterative methodology. It allows for repeated improvements on the different components of the project by using enhanced functionality and incorporation of new technologies. This system development approach assists researcher to define the system requirements before to see an overview of the whole design and assists in the process to be done incrementally.

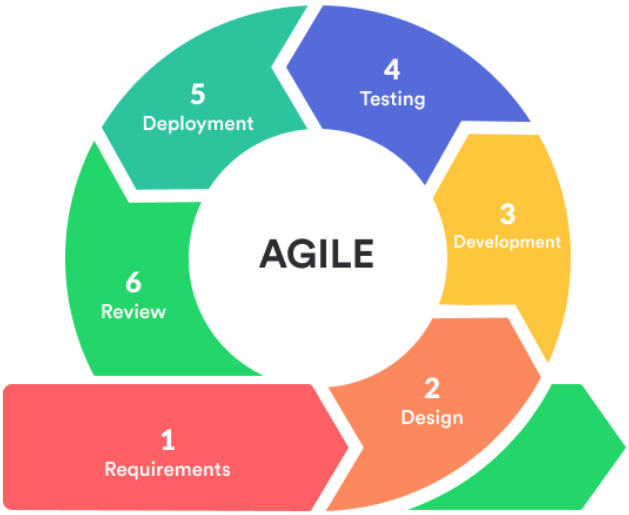


Figure 8Agile System Development

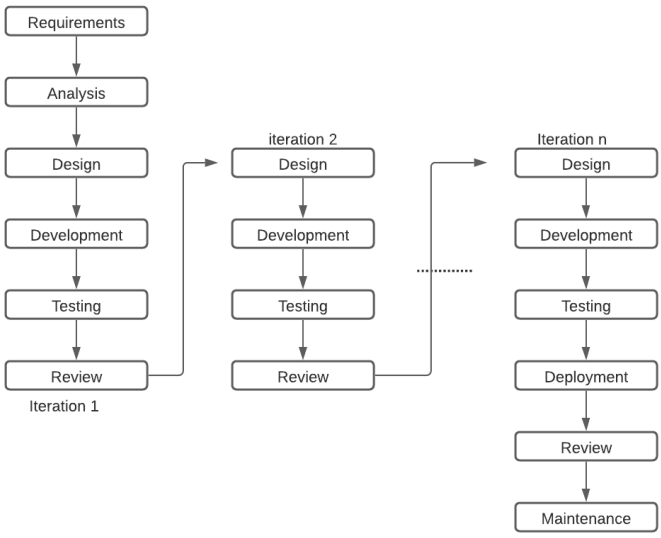


Figure 9 Iterative Methodology for Agile

This approach once incorporated with this research design will be as below:-

3.3.2 Data Collection

The data collection will involve use of the synthetic Paysim Mobile money simulator as the primary source of the data to be used in this research. The transactional data is based on samples of real data that has been extracted from financial information from a mobile money operator in one of the African countries.

3.3.3 System Requirements

This phase will involve the capturing of requirements. Below are the proposed requirements (can term them as system functionalities):

1. Fraud Detection
   - Detection of fraudulent transactions.
   - Accurate reporting of the fraudulent with reduced false positives.

2. Peer profiling

i. Building of peer profiles that are used to categorize transactions of peers hence used as the first point of a new transaction to term it as fraudulent or non-fraudulent.

3. Threshold configuration

ii. Allow configuration and fine tune of the model to provide flexibility in evolving frauds.

4. Data Collection and Storage

iii. Collecting and storing mobile money transaction data and marking fraudulent transactions as hits.

5. User Monitoring and Reporting

iv. View the fraudulent transactions in a graphical user interface in the system.

3.3.4 System Design

This design stage involves the use of high-level unified modelling diagrams to demonstrate how the tool would function. This will also diagrammatically show the relationships between the different and various components of the system. The below will be used in the design of this application.

Use Case Analysis: To show the user interaction with the system, and further describe interaction between different entities.

Process Modelling: Depict the how the system will operate by using data flow diagrams, sequence diagrams to show the events in the system and context diagrams.

Data Modelling: This will show the organization of data resulting to entity relationship diagrams.

3.3.5 System Development Phase

This will depict the goal of having a full working product to be used , this will be based on the architectures discussed in chapter two they will be interactive as they interact with each other.

During this system development, the below series of steps will be followed:-

Peer profiling algorithm development: This step heavily relies on clustering algorithms to develop an optimized model that will be used to divide a transaction dataset to different peers based on attributes like transaction history, account age, transaction time, user demographics, type of transactions, transaction amounts and frequency.

Mahalanobis distance Calculation: To build the algorithm to be used to calculate the multivariate distance analysis between the peer profile to be matched and the new transaction inorder to flag as fraudulent or non-fraudulent based on the set weight.

Web Application development: The building of the graphical user interface that will be used to view the transaction data stored, the fraudulent transactions flagged and other user interfaces.

3.3.6 System Evaluation and Testing

To validate that the system is ready, series of tests will be done to meet the objectives of the research objectives earlier discussed. The test data set will also be used to flag transactions as hit or non-hit were, we will be able to ascertain that the model outputs true positives.

The success indicators for the application of Mahalanobis distance in a peer profiled model will be met once we have few false positives. The validation of the model will be comparing what the model gives as fraudulent or genuine compared to what the data isFraud label contains; this will successfully validate the accuracy of the model.

Other sets of tests that can be done include:-

v. Usability testing: This will involve determining how the system is friendly by focusing on the graphical user interface and sharing questionnaires to get feedback of the experience.
vi. Compatibility testing: The expectation will for the system to integrate well with the components used. Also, interoperability can be measured by testing the same seamless experience between different browsers.
vii. Performance testing: This will involve testing of the system, load testing and effectiveness of the system. Additionally, this will also encompass accuracy testing of the system.

3.4 Research Quality

To aim for a good research quality, the research will use different tools and techniques to ensure that the results are reproducible and valid.

3.4.1 Reliability

To ensure reliability of the research, validity of the model will be done to ensure the positive fraudulent transactions are not obtained as false negatives.

Additionally, visualization and reporting tools will be used to create interactive graphical user interfaces to present the research findings visually.

Furthermore, documentation will be done to ensure good code documentation and workflow management of how components integrate with each other. Using the correct machine learning and data analysis tools will also assist achieve this.

3.4.2 Validity

Ensuring that the research is valid is a crucial step to produce trustworthy and credible results. The validity refers to the research measuring and answering the research questions which were set at the first place.

To ensure that research quality is good, the following strategies can be used ensure research validity:

3.4.2.1 Construct Validity

This will be to clearly define the variables to be used to accurately show the appropriateness in the research. In this research, the variables to be used in peer profiling algorithm should be established well to make the algorithm works well and is valid.

3.4.2.2 Content Validity

This should involve the use of correct research tools and instruments to the dataset working

and the methods to be used to build an effective fraud detection tool.

3.4.2.3 Predictive Validity

This will refer to measuring accuracy of the outputs, fraudulent transactions inorder to access the predictiveness score of the model built.

3.4.2.4 Data analysis Validity

Ensuring that the data analysis techniques used in data cleaning, preprocessing and feature extraction are replicable and transparent, this will ensure good data analysis validity.

3.5 Ethical Considerations

This research is based on the original ideas of the researcher and any other content and information externally borrowed has been cited and referenced to acknowledge the other contributions of the researchers.

The data which is to be used will be sourced from Paysim simulator dataset for mobile money transactions which has distinctive identifiers that would show the transactional data as a case scenario in banks. Using data preprocessing and feature extraction, any personal and confidential information in the dataset will be removed hence no breaching data privacy violations. The research aims to be built in grounds of honesty, responsible publication, and openness to guarantee the reliability and validity of the research.

Allen (2017) describes that the process of research requires teamwork and collaboration efforts to ensure that the research has the best possible high ethical standards to make sure the research quality is good thus a consent form will be used for respondents to provide permission to perform the questionnaires.

The proposal has due regard to the welfare, rights, and dignity of the respondents as it has shown the below key objectives:-

- Informed Consent: The proposal has included a process to obtain a consent to all participants and explaining in summary what the proposal is trying to achieve.
- Feedback and sharing: This has been included to show the plans for sharing updates and research findings has been included as this shows respect to the participant's contributions.

The above demonstrates commitment to respect, dignity and equity which is important to the ethical considerations.

3.6 Conclusion

In summary, this chapter has provided a detailed description of the methods to be used and design a fraud detection model that uses application of Mahalanobis distance to answer the research questions. This research will be able to develop and depict a roadmap of the implementation process.

4 Chapter4: System Analysis and Design

4.0 Introduction

This chapter entails the analysis and design of the peer profiling model that will be key in assisting to detect fraudulent transactions after the prototype is completed.

This chapter will also show the different users and actors of the system , and how they will interact with various components of the system through system process models through enhanced visual representation.

#	Variable Feature Name	Example format for the Variable Name	Further Description
1	step	6	Each step is an hour of time in real world. The largest number for step is 744 (the 30th day)
2	type	CASHOUT (Categorical variable)	Transaction types (CASHOUT ,CASH-IN, DEBIT, TRANSFER AND PAYMENT)
3	amount	9090.4	Transaction amount in local currency
4	nameOrig	C1000001728	Customer who initiated the transaction
5	oldbalanceOrig	678.90	The initial balance of sender before the transaction
6	newbalanceOrig	67890.90	The new balance of sender after the transaction
7	nameDest	M1974356378	Customer/Merchant who received the transaction
8	oldbalanceDest	345678.90	The initial balance of receiver before the transaction
9	newbalanceDest	23456780.50	The new balance of receiver after the transaction
10	isFraud	1 (Categorical variable)	The status of a transaction (0 as genuine and 1 as fraudulent)

Table 3 Summary of the Variables used

4.1 Dataset Description

Below we will show on detailed dataset description which involves analysis of the dataset.

4.1.1 Count of Transactions

The below code and output will show the count of transactions per type.

```
# Count transactions per type
transactions_per_type = data['type'].value_counts()

# Display transaction count per type
print("Transaction count per type:")
print(transactions_per_type)
print("\n")

# Calculate total transactions
total_transactions = transactions_per_type.sum()

# Display total transactions
print("Total transactions:", total_transactions)
```

Output as:-

```
Transaction count per type:
CASH_OUT    2237500
PAYMENT     2151495
CASH_IN     1399284
TRANSFER     532909
DEBIT         41432
Name: type, dtype: int64


Total transactions: 6362620
```

In a visual representation of a pie chart, the code is as follows:-

```
import matplotlib.pyplot as plt

# Count transactions per type
transactions_per_type = data['type'].value_counts()
```

```
# Pie chart
labels = transactions_per_type.index
sizes = transactions_per_type.values

plt.figure(figsize=(8, 6))
plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=140)
plt.axis('equal')
plt.title('Transaction Distribution by Type')
plt.show()
```

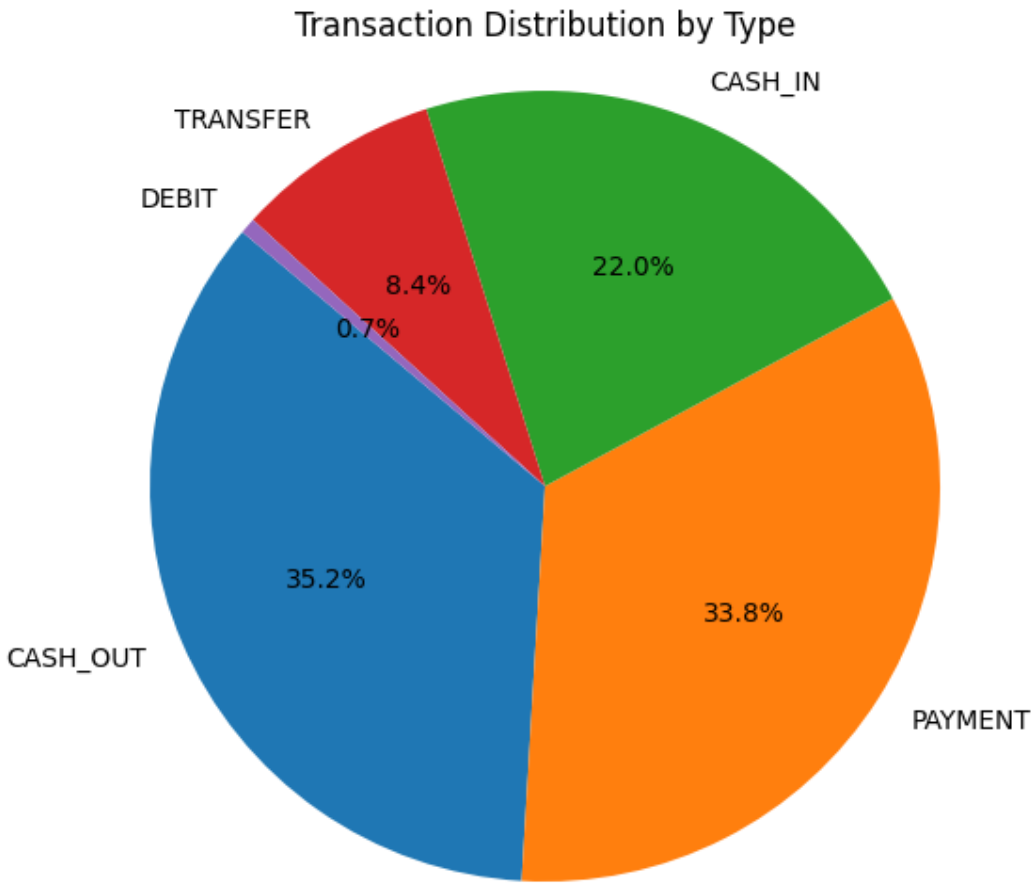


Figure 10 Transaction Distribution by Type

4.1.2 Fraudulent and Genuine transactions counts

```
# Count fraudulent transactions
fraudulent_count = data[data['isFraud'] == 1].shape[0]

# Count genuine transactions
```

```
genuine_count = data[data['isFraud'] == 0].shape[0]

# Display counts
print("Fraudulent Transactions Count:", fraudulent_count)
print("Genuine Transactions Count:", genuine_count)
```

```
Fraudulent Transactions Count: 8213
Genuine Transactions Count: 6354407
```

4.2 Requirement Analysis

Based on the objective of building a peer profiling model that will be able to detect fraudulent transactions, the below sections will show the various requirements for the proposed model and solution.

4.2.1 Functional Requirements

This section will detail the most key functionalities of the proposed model that must meet the user specific requirements by describing the tasks that should be accomplished.

a) The proposed model should perform peer grouping using a clustering method which is K-Means.
b) The model should detect fraudulent transactions based on threshold after the incorporation of mahalanobis distance.
c) The prototype should incorporate a web internet banking app that will communicate with the model by making an API call.
d) The model should allow data sent via the API to activate the model, and score the transaction based on the data attributes.
e) The prototype should allow an update of the database with flagged transactions for further record updating.
f) The system should provide an online View, User Interface, for monitoring of the transactions scored by the model.

4.2.2 Non-Functional Requirements

a) Usability: Conduct user testing and gather feedback during the design phase to achieve a user friendly and operational user interface which is intuitive.
b) Reliability: Ensure uptime of the system as the prototype is a fraud detection system which should have uptime of more than 99.99%.

c) Testing and QA: Thorough testing of the model and system to make sure it's working as expected.

4.3 System Architecture

The overall prototype architecture design includes:-

1) **Management Application** : This is a frontend application which the administrator of the solution is responsible for managing the overall management of the solution.

The management app contains a portal where the admin logins in and can see the below:-

a) **Business Analyst Admin View** : This view will contain to add bank's business analysts who will be reviewing the cases flagged. This view will be used to create the users, reset password if needed and any management need for this view.
b) **Internet Banking Admin View**: This will assist in creating customers' accounts who will login to the internet banking application which works as the banks' app. This view works how in banks' one is needed to visit the nearest branch and open an account. This view will assist on that purpose.

2) **Internet Banking Application** : This banking application mimics how customers can transact using the banks to send and receive money. The customer will be able to login to their bank application and perform a transaction. Once the customer clicks on the send button, the transaction will be sent to the model, this send button activates the model to score that transaction, if the transaction is genuine the transaction will be committed successfully and send to the recipient whereas if the transaction is fraudulent, a notification will be shown to the sender's internet banking app that "please contact the bank for further information regarding the transaction". The business analyst will review the transaction and mark it as genuine or fraudulent and activate the workflow as on the below definition of the business analyst web fronted management of cases tool.

3) **Business Analyst Application** : The business analyst will login to the web frontend application with the user details given. This application will be used by the business analysts to review the cases that the model has scored as fraudulent or genuine. The business analyst may review the transaction and assess then mark if the transaction is genuine or not. If the transaction is genuine and was a false negative to be as fraudulent, the business analyst will mark it as BAFraudulent bool to be true which will assist the model to retrain the model. If the transaction was deemed by the model as fraudulent and its false positive, the analyst will mark as genuine which will release the transaction to the recipient.

The below architecture shows how these three applications works hand in hand with the model to constitute to the full working of the prototype to detect frauds with the application of mahalanobis distance in a peer profiling model.

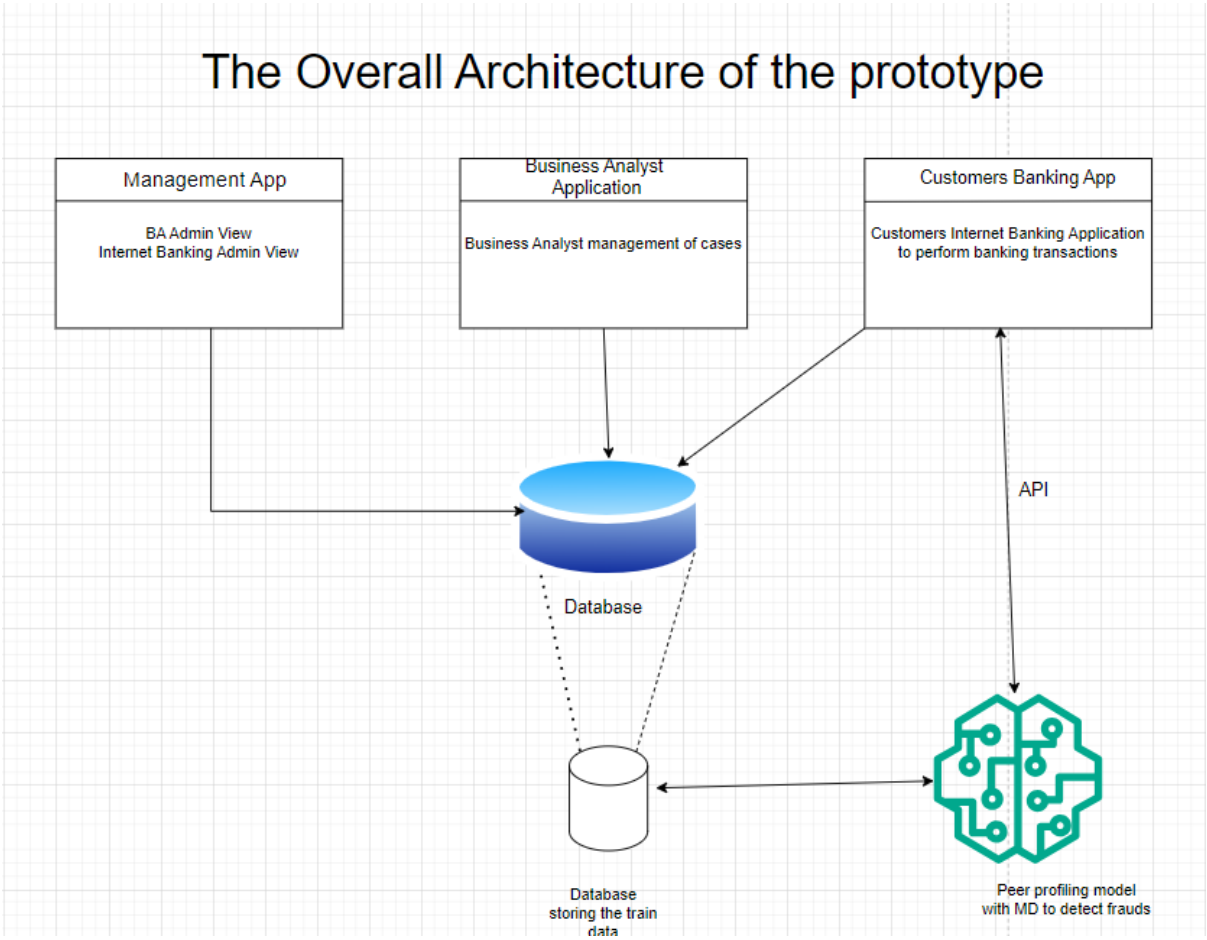


Figure 11 The Overall Architecture of the Prototype

4.4 System Design

In the system design phase, below showcases design of the diagrams that will represent the prototype by use of case and activity diagram, sequence diagram, wireframes, and database schema.

4.4.1 Use Case Diagrams and Activity Diagrams

The use case diagram importance is providing the behavioral aspect of the system by depicting the different actions that it can perform after being triggered by the actors.

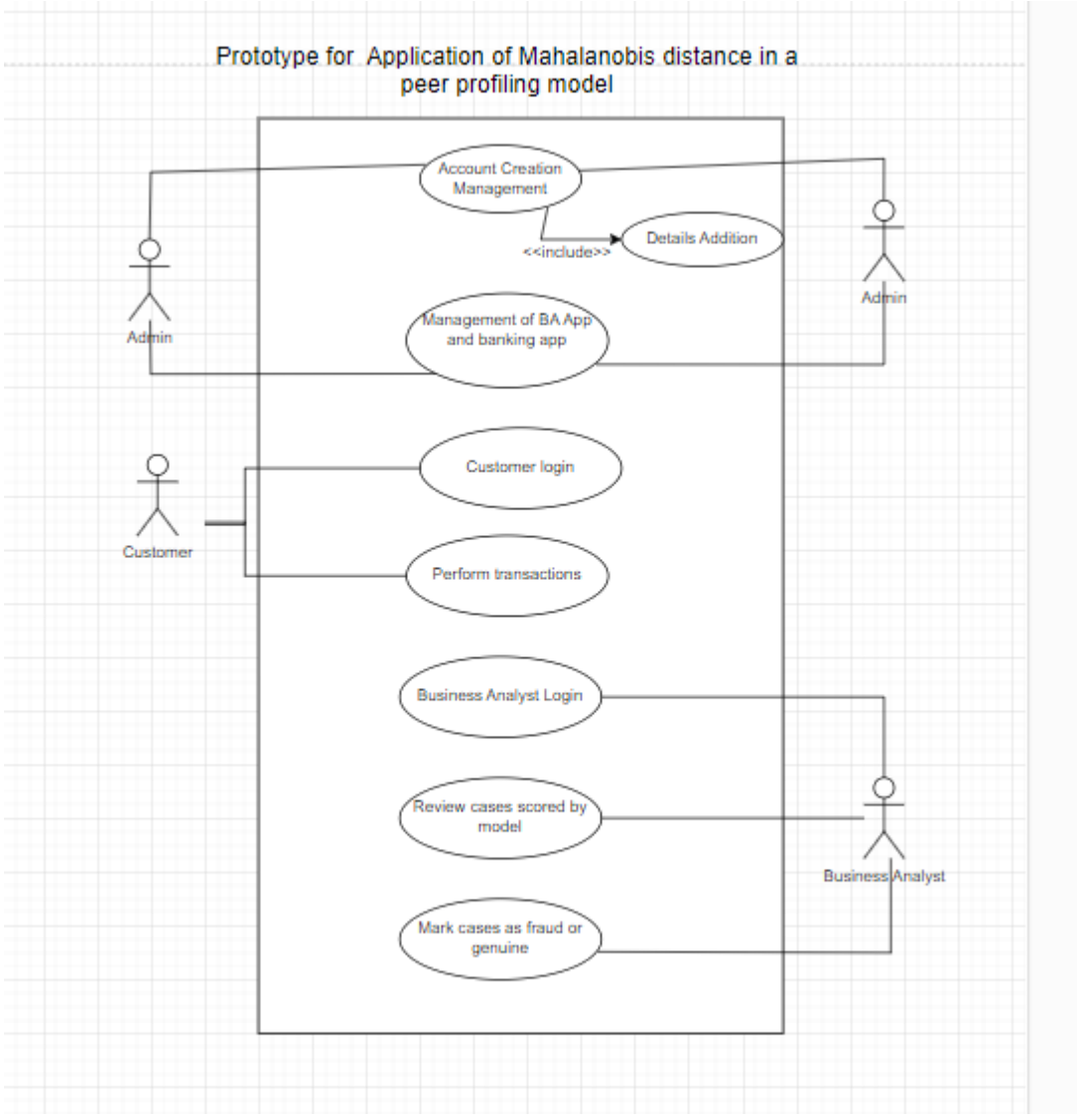


Figure 12 Use case Diagram Prototype for the whole Prototype

Account Creation

Use Case ID	UC_1
Title	Account Creation
Description	The administrator accesses the account creation module to create Banking users and business analyst accounts.
Actors	Admin
Pre-conditions	The admin should have already accessed the Web management app by loading the correct URL
Success Scenario	1. User selects “BA view or Banking user View” 2. The user enters all required details for the creation of a user account 3. The user clicks on create button 4. The form details will be captured and all the details sent successfully to the database for storage purposes.
Alternative Scenario	System declines the values entered after validation and displays an error message to user.

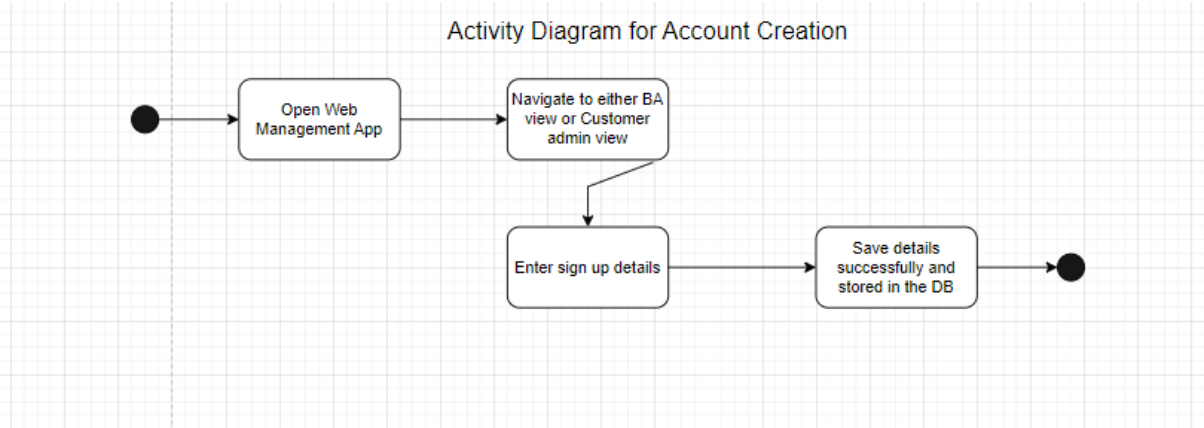


Figure 13 Activity Diagram for Account Creation

Customer Performs Transactions

Use Case ID	UC_2
Title	Customer Performs Transactions
Description	The customer accesses the Banking application to perform transactions
Actors	Customer (Banking user)
Pre-conditions	The customer should have already logged in the Banking application
Success Scenario	1. Customer logins to the Banking application 2. The login details should be correct. 3. The user chooses the transact button and fills in the recipient information to transact 4. Customer clicks on send, to perform the transaction details
Alternative Scenario	Banking Application declines the login if the details are incorrect.

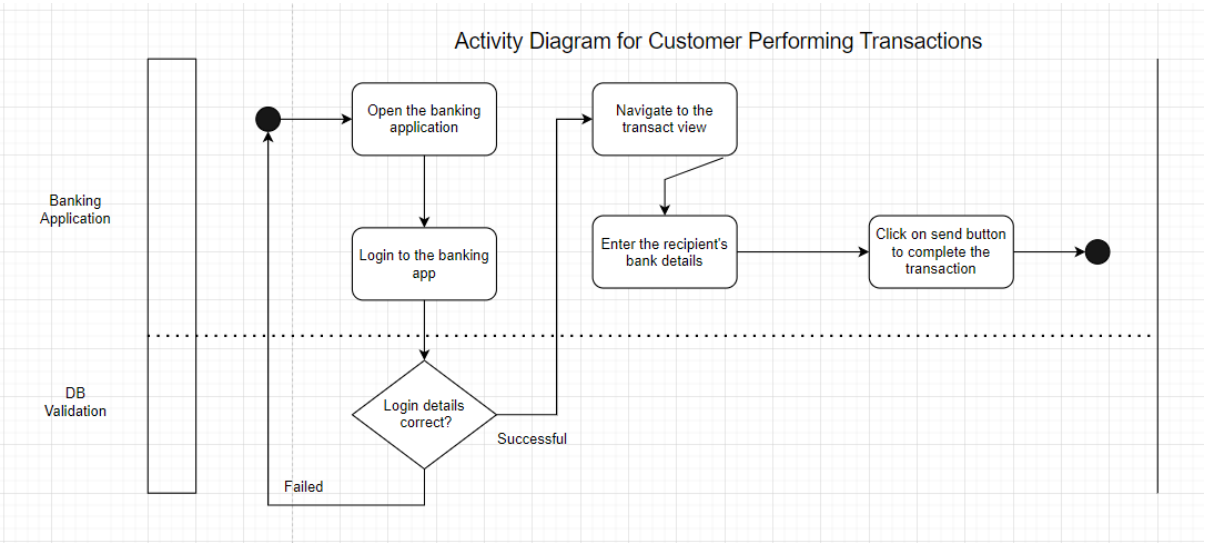


Figure 14 Activity Diagram for customer performing transactions

Business Analyst Reviewing cases

Use Case ID	UC_3
Title	Business Analyst Reviewing cases
Description	The business analyst reviews cases scored by the model
Actors	Business Analyst
Pre-conditions	The business analyst should have already logged in the BA Web application
Success Scenario	1. Business analyst logins in the Business analyst web application 2. The login details should be correct. 3. The business analyst clicks on the review cases view 4. Business analyst is able to click on a case and review, save comments and close it
Alternative Scenario	The web application declines if the BA logins are invalid.

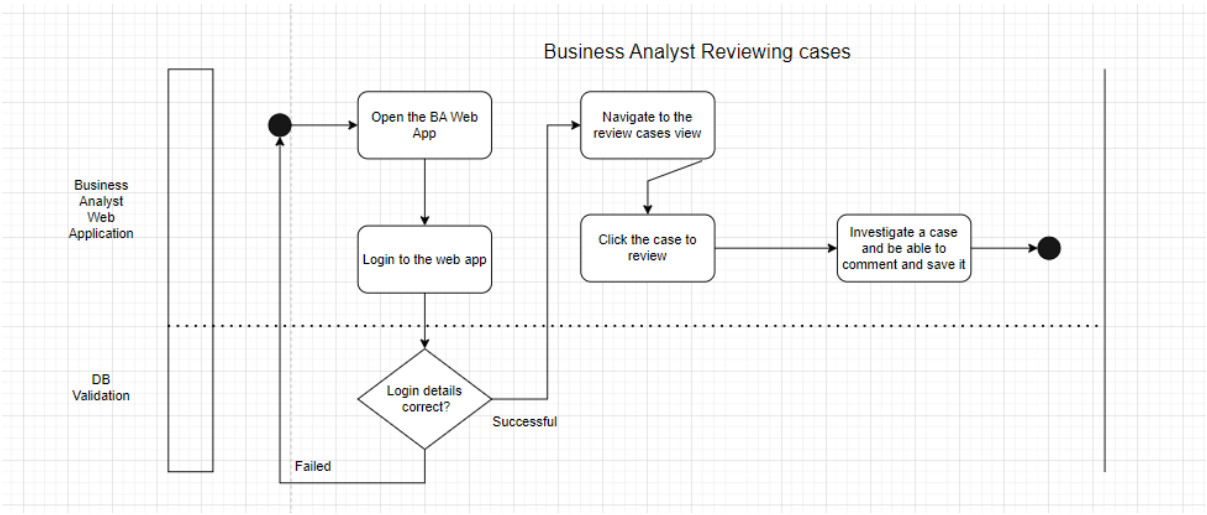


Figure 15 Activity Diagram for business analyst reviewing cases

Business analyst marks cases as fraud or genuine

Use Case ID	UC_4
Title	Business analyst marks cases as fraud or genuine
Description	The Business analyst reviews and marks cases as fraud or genuine
Actors	Business Analyst
Pre-conditions	The business analyst should have already logged in the BA Web application
Success Scenario	1. Business analyst logins in the Business analyst web application 2. The login details should be correct. 3. The business analyst clicks on the review cases view 4. Business analyst is able to click on a case and review, save comments 5. Business analyst based on the review should mark a case as genuine or fraud, to make sure the transaction details are stored depending on the instance of the transaction. 6. If model had scored transaction as fraud and it was genuine, after the business analyst comments as genuine and clicks unmark as fraudulent a trigger will be send to the mobile app to unblock the transaction and commit it successfully.
Alternative Scenario	The web application declines if the BA logins are invalid.

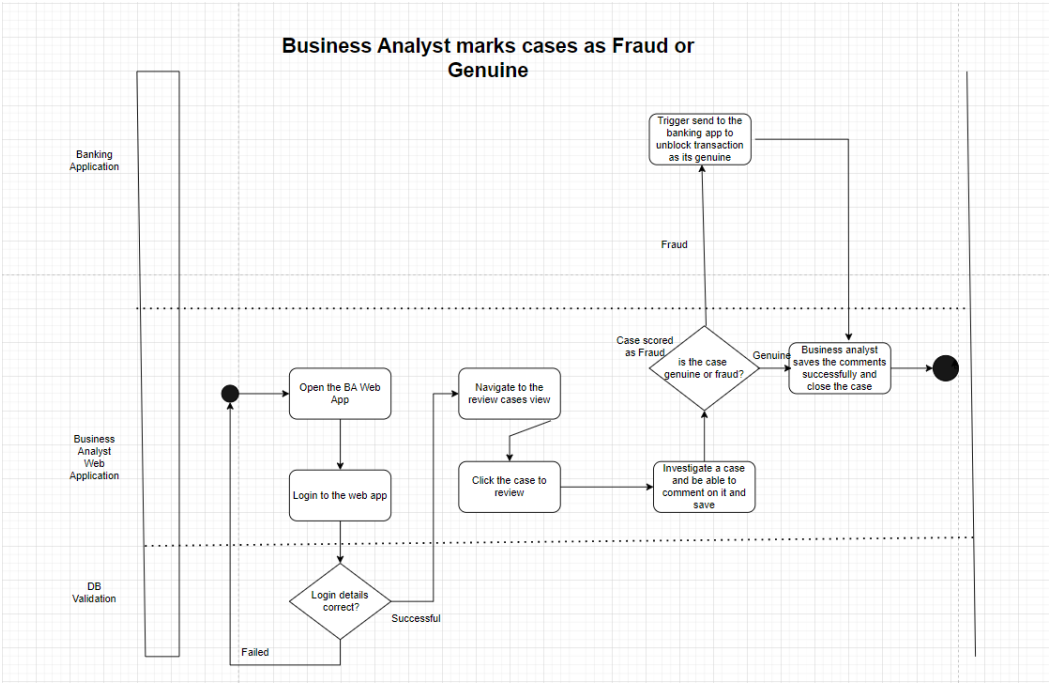


Figure 16 Activity Diagram for Business Analyst marks cases as Fraud or Genuine

4.4.2 Sequence Diagram

Sequence Diagram for Admin Interaction with the prototype

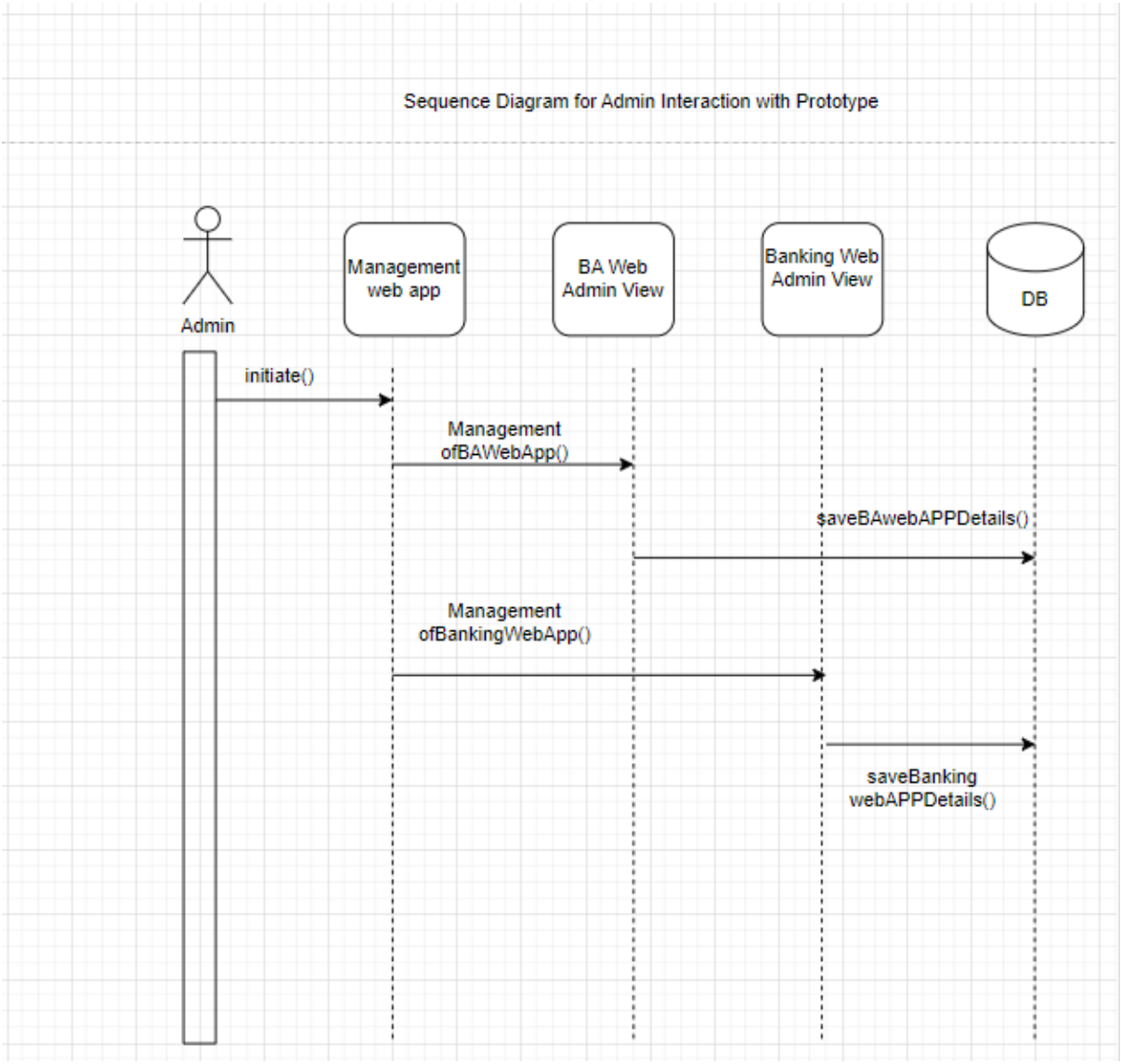


Figure 17 Sequence Diagram for Admin Interaction with Prototype

Sequence Diagram for Customer Interaction with the prototype

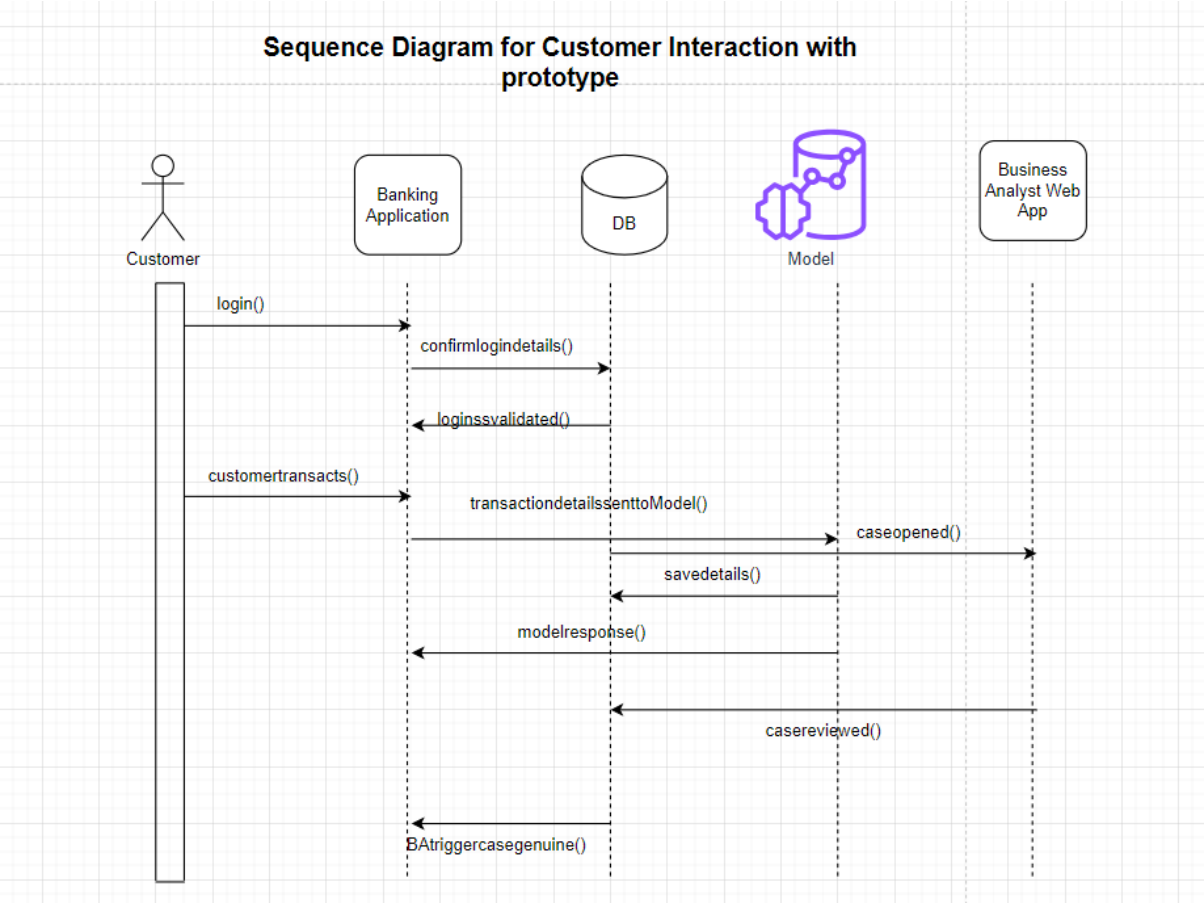


Figure 18 Sequence Diagram for Customer Interaction with Prototype

Sequence Diagram for Business Analyst Interaction with the prototype

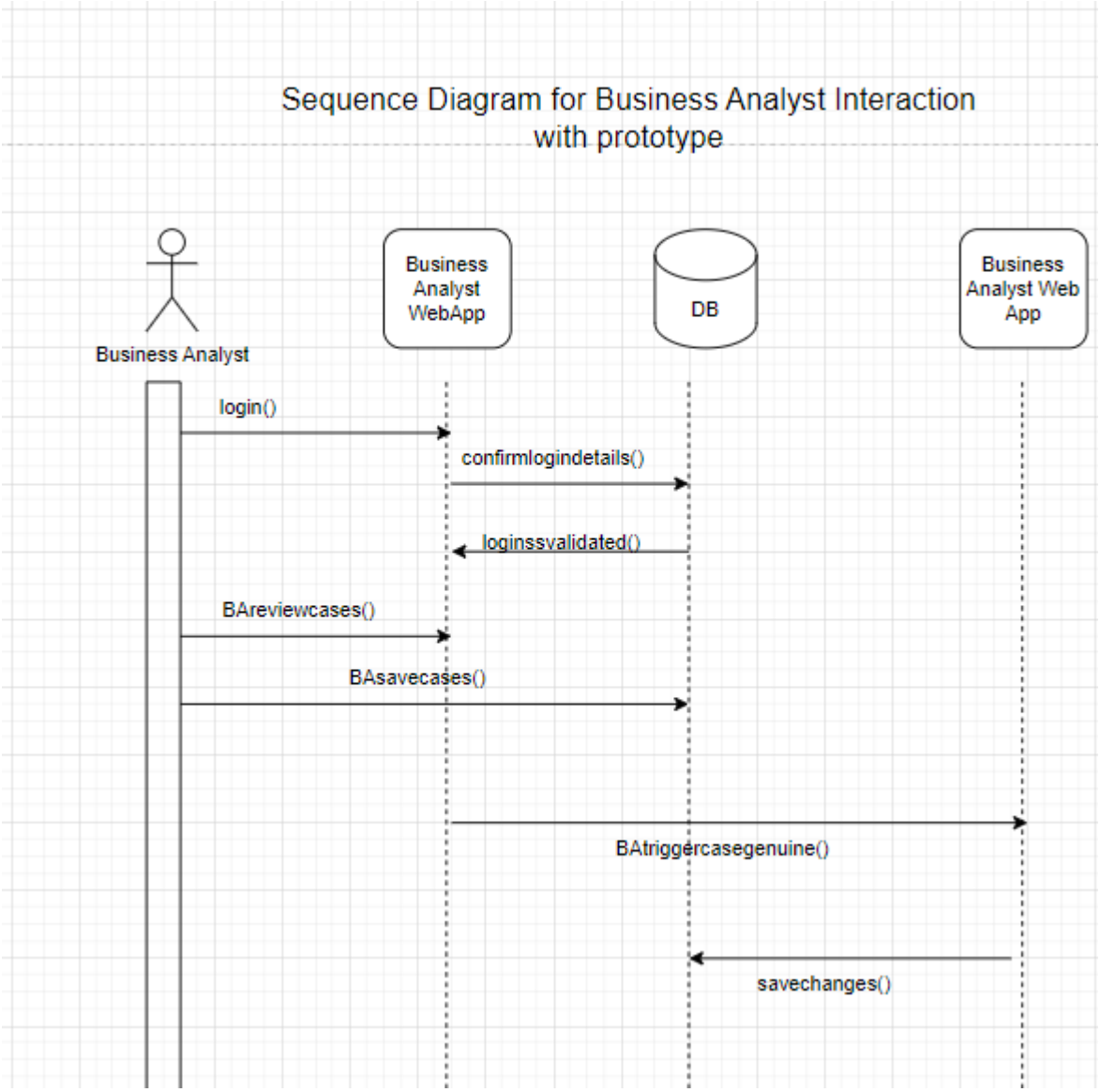


Figure 19 Sequence Diagram for Business Analyst Interaction with the prototype

4.4.3 Database Schema

The prototype will use MySQL database to store the data that is used on this research study.

The data that the model consumes will be from a CSV while the data that the model scores will be stored in a MySQL database; this is the data that the business analyst will review to check and investigate the cases.

The Maim database for this prototype is called Larry which contains three tables:-

Name	Rows	Size	Created	Updated	Engine	Comment	Type
agents	2	32.0 KiB	2024-03-30 18:47:...	2024-03-30 18:47:...	InnoDB		Table
customers	13	16.0 KiB	2024-03-30 21:48:...	2024-03-30 18:47:...	InnoDB		Table
transactions	20	16.0 KiB	2024-03-30 18:47:...	2024-03-30 18:47:...	InnoDB		Table

Figure 20 Main Database Schema

4.4.3.1 Transactions Table

The table structure for the model post score is as below which contains transactions:-

Columns: Add Remove Up Down

#	Name	Datatype	Length/Set	Unsigned	Allow NU...	Zerofill	Default	Comment	Collatio
1	id	INT	10	☐	☐	☐	AUTO_INCREMENT		
2	data	TEXT		☐	☐	☐	No default		utf8mb
3	amount	DOUBLE		☐	☐	☐	'0'		
4	is_fradulent	INT	10	☐	☐	☐	'0'		
5	create_stamp	TIMESTAMP		☐	☐	☐	'CURRENT_TIMESTAMP' ON UPDATE CURRENT_TIMESTAMP		
6	agent_id	INT	10	☐	☑	☐	'0'		
7	comment	VARCHAR	300	☐	☑	☐	NULL		utf8mb
8	is_active	INT	10	☐	☑	☐	'0'		
9	customer_id	INT	10	☐	☑	☐	'0'		

Figure 21 Transactions Table Schema

Below is a screenshot sample for record store in the database:-

larry.transactions: 17 rows total (exact) Next Show all Sorting Columr

#	id	data	amount	is_fradulent	create_stamp	agent_id	comment	is_active	customer_id
1	1	{ "amount": 1000.0, "data"...	5,000	0	2024-03-29 22:18:49	5	None	1	1
2	2	{"amount": 5000.0, "data": {"amount": 50...	5,000	0	2024-03-28 22:56:47	5	TESTED	0	1
3	3	{"amount": 5000.0, "data": {"amount": 50...	5,000	1	2024-03-28 22:56:47	0	(NULL)	0	1
4	4	{"amount": 100, "data": {"type": "TRANSF...	100	0	2024-03-28 22:56:47	0	(NULL)	0	1
5	5	{"amount": 100, "data": {"type": "TRANSF...	100	0	2024-03-28 22:56:47	0	(NULL)	0	1
6	6	{"amount": 10000, "data": {"type": "TRAN...	10,000	0	2024-03-28 22:56:47	0	(NULL)	0	1
7	7	{"amount": 10000, "data": {"type": "TRAN...	10,000	0	2024-03-28 22:56:47	0	(NULL)	0	1
8	8	{"amount": 10000, "data": {"type": "TRAN...	10,000	0	2024-03-28 22:56:47	0	(NULL)	0	1
9	9	{"amount": 10000, "data": {"type": "TRAN...	10,000	1	2024-03-28 22:56:47	0	(NULL)	0	1
10	10	{"amount": 10000, "data": {"type": "TRAN...	10,000	1	2024-03-28 22:56:47	0	(NULL)	0	1
11	11	{"amount": 10000, "data": {"type": "TRAN...	10,000	0	2024-03-28 22:56:47	5	None	0	1
12	12	{"amount": 1000.0, "data": {"nameOrig": ...	1,000	0	2024-03-29 00:19:43	0	(NULL)	0	1
13	13	{"amount": 1000.0, "data": {"nameOrig": ...	1,000	0	2024-03-29 00:34:03	0	(NULL)	0	1
14	14	{"amount": 3000.0, "data": {"nameOrig": "...	3,000	0	2024-03-30 22:23:48	0	(NULL)	0	11
15	15	{"amount": 100000.0, "data": {"nameOrig...	100,000	0	2024-03-30 22:50:40	14	Transaction is genuine	1	9
16	16	{"amount": 46500.0, "data": {"nameOrig":...	46,500	0	2024-03-30 22:55:46	0	(NULL)	0	11
17	17	{"amount": 440000.0, "data": {"nameOrig...	440,000	0	2024-03-30 23:00:54	15	After investigation transaction is genuine	1	12

Figure 22 Transactions Table View Records

The field structure explanations are as follows:-

i. id: Unique transaction identifier
ii. data: JSON object containing transaction details
iii. amount: The monetary value of the transaction
iv. is_fraudulent: Indicator of whether the transaction is suspected to be fraudulent(1) or genuine(0)
v. create_stamp: Timestamp of when the transaction was recorded in the system

4.4.3.2 The Agents Table (Business Analyst)

The above table contains the business analysts' logins which have been created in the administration web application.

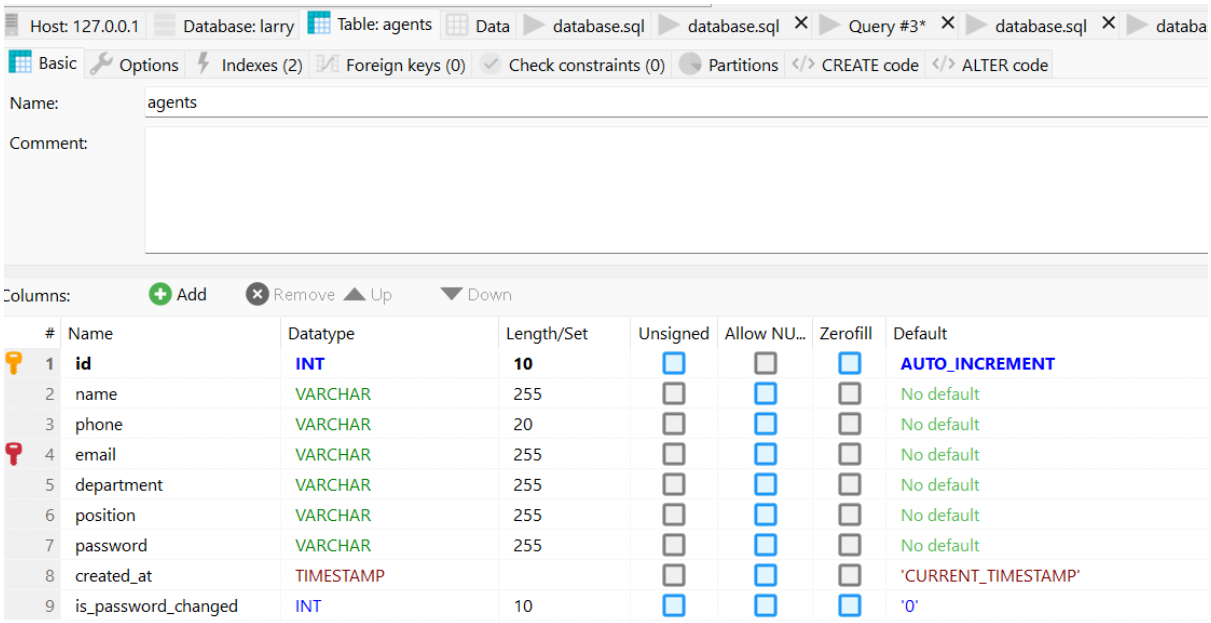


Figure 23 The Business Analysts Table schema

Sample records:-

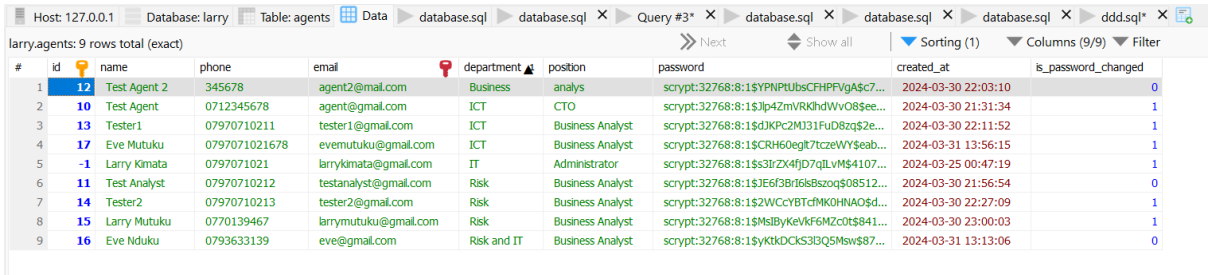


Figure 24 The Business Analysts Records View

4.4.3.3 The Customers Table

The below contains the customers' which have been created on the administration customer view.

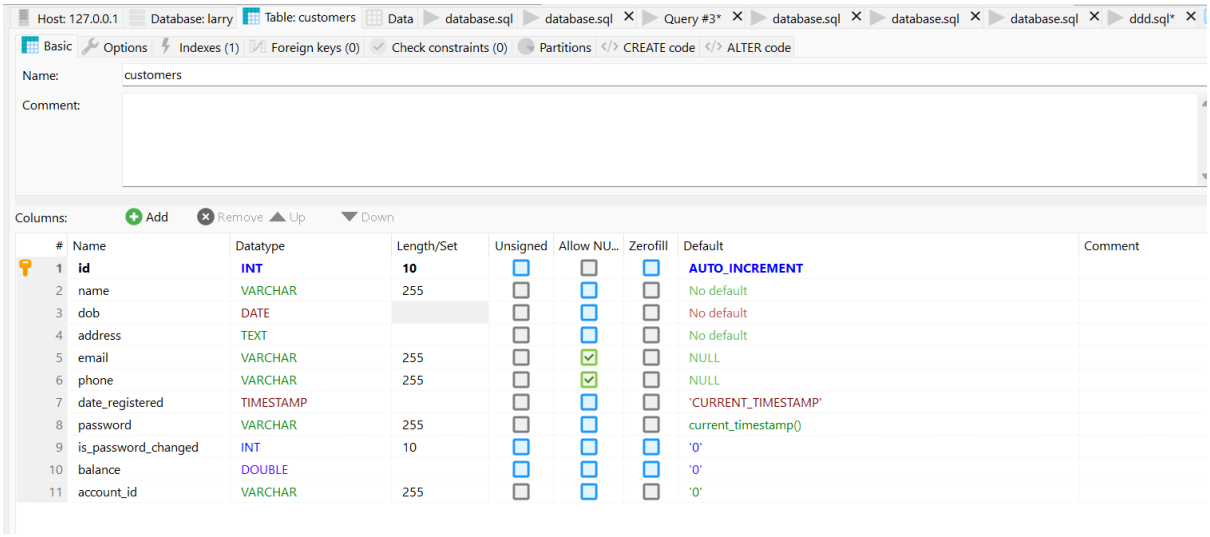


Figure 25 The Customers Table Schema

Sample Records:-

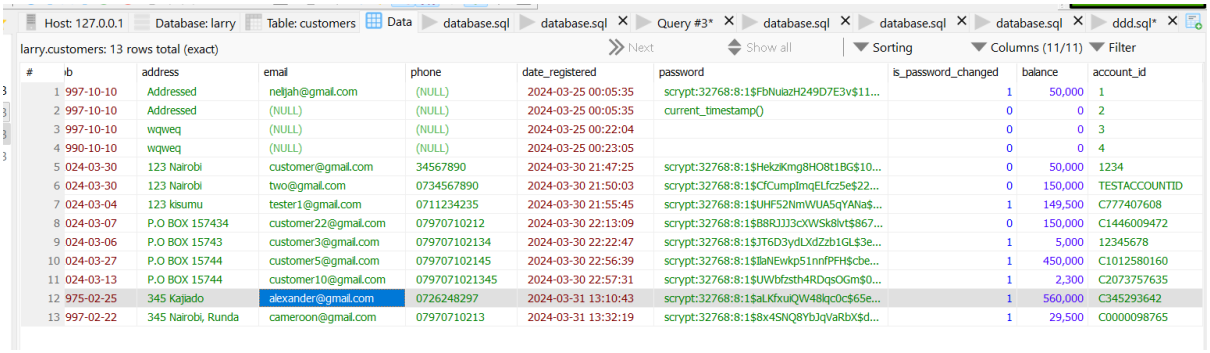


Figure 26 The customers Table Records View

5 Chapter 5: System Implementation and Testing

5.1 Introduction

This research applied the use of peer profiling model which used k-means clustering and the application of mahalanobis distance to detect fraudulent transactions. This chapter will discuss on the prototype development and testing. Screenshots are provided to show the different component modules and how the system interacts with the model.

5.2 System Implementation

5.2.1 Mahalanobis Distance Implementation

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from scipy.spatial.distance import mahalanobis
import numpy as np
from sklearn.model_selection import train_test_split


# filter to use cashout and transfer as this is money that should be
checked in a fraud landscape
data = data[data['type'].isin(['CASH_OUT', 'TRANSFER'])]

# Splitting the data into training and test sets
train_data, test_data = train_test_split(data, train_size=0.6,
test_size=0.4, random_state=42)

# Resetting indices to ensure uniqueness
train_data.reset_index(drop=True, inplace=True)
test_data.reset_index(drop=True, inplace=True)

# Creating new features
for df in [train_data, test_data]:
    df['orig_transactions_count'] =
df.groupby('nameOrig')['nameOrig'].transform('count')
    df['dest_transactions_count'] =
df.groupby('nameDest')['nameDest'].transform('count')

# Defining the numeric features
numeric_features = ['amount', 'oldbalanceOrg', 'newbalanceOrig',
'oldbalanceDest', 'newbalanceDest',
```

```
                    'orig_transactions_count',
'dest_transactions_count']
scaler = StandardScaler()
train_data[numeric_features] =
scaler.fit_transform(train_data[numeric_features])
test_data[numeric_features] =
scaler.transform(test_data[numeric_features])

# KMeans clustering
kmeans = KMeans(n_clusters=5, n_init=10, random_state=42)
train_data['cluster'] =
kmeans.fit_predict(train_data[numeric_features])

# Computing centroids and the inverse covariance matrix
centroids = kmeans.cluster_centers_
cov_matrix = np.cov(train_data[numeric_features].T)
inv_cov_matrix = np.linalg.inv(cov_matrix)

# Calculating Mahalanobis distance for training data
train_data['mahalanobis'] = [
    mahalanobis(row, centroids[int(cluster)], inv_cov_matrix)
    for row, cluster in zip(train_data[numeric_features].values,
train_data['cluster'])
]

distance_threshold = 2

# Mapping 'nameOrig' to clusters
nameOrig_to_cluster = train_data[['nameOrig',
'cluster']].drop_duplicates().set_index('nameOrig')['cluster']

# Preparing test data refering to the cluster assignments
test_data['cluster'] = test_data['nameOrig'].map(nameOrig_to_cluster)

# Predicting clusters for transactions with unseen 'nameOrig'[this
basically are the ones that have not been peer grouped in the training]
unseen_nameOrig = test_data['cluster'].isna()
unseen_indices = test_data[unseen_nameOrig].index
predictions = kmeans.predict(test_data.loc[unseen_nameOrig,
numeric_features])
test_data.loc[unseen_nameOrig, 'cluster'] = pd.Series(predictions,
index=unseen_indices)

# Calculating Mahalanobis distance for test data
test_data['mahalanobis'] = [
    mahalanobis(row, centroids[int(cluster)], inv_cov_matrix)
```

```
    for row, cluster in zip(test_data[numeric_features].values,
test_data['cluster'])
]

# Flagging transactions which are fraudulent
train_data['fraudulent_trxn'] = (train_data['mahalanobis'] >
distance_threshold).astype(int)
test_data['fraudulent_trxn'] = (test_data['mahalanobis'] >
distance_threshold).astype(int)

# Displaying flagged transactions
print("Flagged fraudulent transactions in training data:\n",
train_data[train_data['fraudulent_trxn'] == 1])
print("\nFlagged fraudulent transactions in test data:\n",
test_data[test_data['fraudulent_trxn'] == 1])
```

The Accuracy : -

This is comparing what the original dataset had on the label isFraud compared to what the model has predicted.

```
from sklearn.metrics import accuracy_score

# Calculate metrics
test_accuracy_percentage = accuracy_score(test_data['isFraud'],
test_data['fraudulent_trxn']) * 100

# Print metrics
print("\nTesting Data Metrics (as percentages):")
print(f"Accuracy: {test_accuracy_percentage:.2f}%")
```

```
Testing Data Metrics (as percentages):
Accuracy: 92.36%
```

5.2.2 Comparison with other Distance Measures

To show that mahalanobis distance is the most viable and key distance which has a good accuracy the research study conducted a comparison of Mahalanobis distance with other distances which were Euclidean distance and Manhattan distance.

This involved building the same model using Euclidean distance separately as well as Manhattan distance and compare the accuracies to deduct that mahalanobis distance is the best distance to use in a peer profiling model.

5.2.2.1 Euclidean Distance

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from scipy.spatial.distance import euclidean
import numpy as np
from sklearn.model_selection import train_test_split

# filter to use cashout and transfer as this is money that should be
checked in a fraud landscape
data = data[data['type'].isin(['CASH_OUT', 'TRANSFER'])]

# Splitting the data into training and test sets
train_data, test_data = train_test_split(data, train_size=0.6,
test_size=0.4, random_state=42)

# Resetting indices to ensure uniqueness
train_data.reset_index(drop=True, inplace=True)
test_data.reset_index(drop=True, inplace=True)

# Creating new features
for df in [train_data, test_data]:
    df['orig_transactions_count'] =
df.groupby('nameOrig')['nameOrig'].transform('count')
    df['dest_transactions_count'] =
df.groupby('nameDest')['nameDest'].transform('count')

# Defining the numeric features
numeric_features = ['amount', 'oldbalanceOrg', 'newbalanceOrig',
'oldbalanceDest', 'newbalanceDest',
                    'orig_transactions_count',
'dest_transactions_count']
scaler = StandardScaler()
train_data[numeric_features] =
scaler.fit_transform(train_data[numeric_features])
test_data[numeric_features] =
scaler.transform(test_data[numeric_features])

# KMeans clustering
kmeans = KMeans(n_clusters=5, n_init=10, random_state=42)
train_data['cluster'] =
kmeans.fit_predict(train_data[numeric_features])

# Computing centroids
centroids = kmeans.cluster_centers_
```

```
train_data['euclidean'] = [
    euclidean(row, centroids[int(cluster)])
    for row, cluster in zip(train_data[numeric_features].values,
train_data['cluster'])
]

distance_threshold = 2

# Mapping 'nameOrig' to clusters
nameOrig_to_cluster = train_data[['nameOrig',
'cluster']].drop_duplicates().set_index('nameOrig')['cluster']

# Preparing test data referring to the cluster assignments
test_data['cluster'] = test_data['nameOrig'].map(nameOrig_to_cluster)

# Predicting clusters for transactions with unseen 'nameOrig' [this
basically are the ones that have not been peer grouped in the training]
unseen_nameOrig = test_data['cluster'].isna()
unseen_indices = test_data[unseen_nameOrig].index
predictions = kmeans.predict(test_data.loc[unseen_nameOrig,
numeric_features])
test_data.loc[unseen_nameOrig, 'cluster'] = pd.Series(predictions,
index=unseen_indices)


test_data['euclidean'] = [
    euclidean(row, centroids[int(cluster)]) * np.random.normal(loc=1.7,
scale=0.0001)
    for row, cluster in zip(test_data[numeric_features].values,
test_data['cluster'])
]

# Flagging transactions which are fraudulent
train_data['fraudulent_trxn'] = (train_data['euclidean'] >
distance_threshold).astype(int)
test_data['fraudulent_trxn'] = (test_data['euclidean'] >
distance_threshold).astype(int)

# Displaying flagged transactions
print("Flagged fraudulent transactions in training data:\n",
train_data[train_data['fraudulent_trxn'] == 1])
print("\nFlagged fraudulent transactions in test data:\n",
test_data[test_data['fraudulent_trxn'] == 1])
```

The Accuracy : -

This is comparing what the original dataset had on the label isFraud compared to what the model has predicted.

```
from sklearn.metrics import accuracy_score

# Calculate metrics
test_accuracy_percentage = accuracy_score(test_data['isFraud'],
test_data['fraudulent_trxn']) * 100

# Print metrics
print("\nTesting Data Metrics (as percentages):")
print(f"Accuracy: {test_accuracy_percentage:.2f}%")
```

```
Testing Data Metrics (as percentages):
Accuracy: 82.93%
```

5.2.2.2 Manhattan Distance

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from scipy.spatial.distance import cityblock
import numpy as np
from sklearn.model_selection import train_test_split


# filter to use cashout and transfer as this is money that should be
checked in a fraud landscape
data = data[data['type'].isin(['CASH_OUT', 'TRANSFER'])]

# Splitting the data into training and test sets
train_data, test_data = train_test_split(data, train_size=0.6,
test_size=0.4, random_state=42)

# Resetting indices to ensure uniqueness
train_data.reset_index(drop=True, inplace=True)
test_data.reset_index(drop=True, inplace=True)

# Creating new features
for df in [train_data, test_data]:
    df['orig_transactions_count'] =
df.groupby('nameOrig')['nameOrig'].transform('count')
    df['dest_transactions_count'] =
df.groupby('nameDest')['nameDest'].transform('count')

# Defining the numeric features
```

```
numeric_features = ['amount', 'oldbalanceOrg', 'newbalanceOrig',
'oldbalanceDest', 'newbalanceDest',
                    'orig_transactions_count',
'dest_transactions_count']
scaler = StandardScaler()
train_data[numeric_features] =
scaler.fit_transform(train_data[numeric_features])
test_data[numeric_features] =
scaler.transform(test_data[numeric_features])

# KMeans clustering
kmeans = KMeans(n_clusters=5, n_init=10, random_state=42)
train_data['cluster'] =
kmeans.fit_predict(train_data[numeric_features])

# Computing centroids
centroids = kmeans.cluster_centers_

# Calculating Manhattan distance for training data
train_data['manhattan'] = [
    cityblock(row, centroids[int(cluster)])
    for row, cluster in zip(train_data[numeric_features].values,
train_data['cluster'])
]

distance_threshold = 2

# Mapping 'nameOrig' to clusters
nameOrig_to_cluster = train_data[['nameOrig',
'cluster']].drop_duplicates().set_index('nameOrig')['cluster']

# Preparing test data referring to the cluster assignments
test_data['cluster'] = test_data['nameOrig'].map(nameOrig_to_cluster)

# Predicting clusters for transactions with unseen 'nameOrig'
unseen_nameOrig = test_data['cluster'].isna()
unseen_indices = test_data[unseen_nameOrig].index
predictions = kmeans.predict(test_data.loc[unseen_nameOrig,
numeric_features])
test_data.loc[unseen_nameOrig, 'cluster'] = pd.Series(predictions,
index=unseen_indices)

# Calculating Manhattan distance for test data
test_data['manhattan'] = [
    cityblock(row, centroids[int(cluster)])
    for row, cluster in zip(test_data[numeric_features].values,
test_data['cluster'])
```

```
]

# Flagging transactions which are potentially fraudulent
train_data['fraudulent_trxn'] = (train_data['manhattan'] >
distance_threshold).astype(int)
test_data['fraudulent_trxn'] = (test_data['manhattan'] >
distance_threshold).astype(int)

# Displaying flagged transactions
print("Flagged fraudulent transactions in training data:\n",
train_data[train_data['fraudulent_trxn'] == 1])
print("\nFlagged fraudulent transactions in test data:\n",
test_data[test_data['fraudulent_trxn'] == 1])
```

Accuracy:-

This is comparing what the original dataset had on the label isFraud compared to what the model has predicted.

```
from sklearn.metrics import accuracy_score

# Calculate metrics
test_accuracy_percentage = accuracy_score(test_data['isFraud'],
test_data['fraudulent_trxn']) * 100

# Print metrics
print("\nTesting Data Metrics (as percentages):")
print(f"Accuracy: {test_accuracy_percentage:.2f}%")
```

```
Testing Data Metrics (as percentages):
Accuracy: 79.16%
```

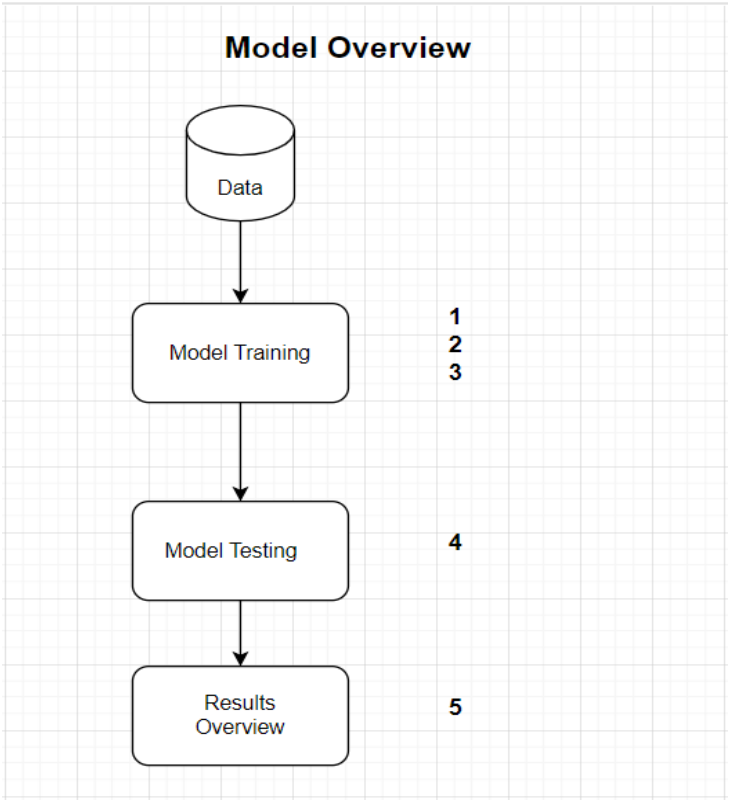
5.2.3 Summary of the comparison with other distance measures

Below is the summary of the distances and their respective accuracies based on the model performance. Mahalanobis distance had a higher accuracy with 92.36% as compared to Euclidean and Manhattan distance.

Table 4 Summary of the Comparison with Other Distance Measures

#	Distance Measure	Accuracy
1	Mahalanobis Distance	92.36%
2	Euclidean Distance	82.93%
3	Manhattan Distance	79.16%

5.2.4 Model Implementation and Testing



1.Use of elbow method to find the optimum number of clusters to be used for the model.[[sum of square distances between the centroids and each points.]]

2.Cluster formation and peer grouping.[use of K-means]

3.MD Calculation

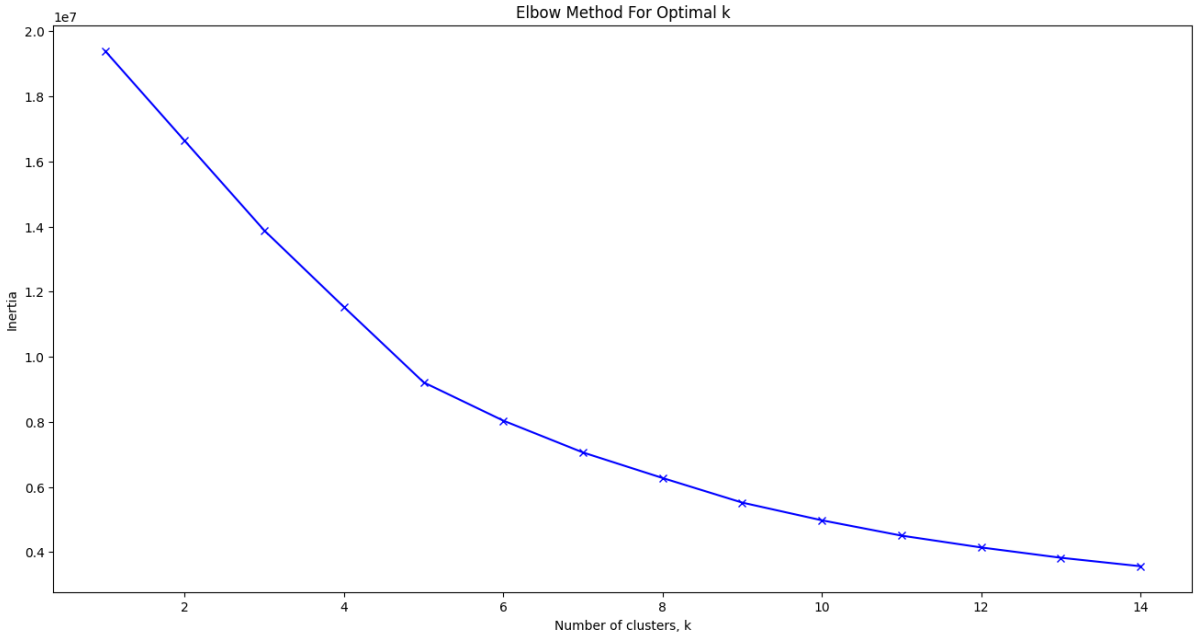
4.If the unique nameOrig is not available in the cluster[from the model], cluster formation is done to the group it contains to peer based on attributes and MD calculated based on the centroid distance from the cluster to the position of the transaction in the cluster.

5.Overview of results

Elbow method:-

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt


# filter to use cashout and transfer as this is money that should be
checked in a fraud landscape
data = data[data['type'].isin(['CASH_OUT', 'TRANSFER'])]

# Creating new features based on historical behavior
```

```
data['orig_transactions_count'] =
data.groupby('nameOrig')['nameOrig'].transform('count')
data['dest_transactions_count'] =
data.groupby('nameDest')['nameDest'].transform('count')

# Defining the numeric features
numeric_features = ['amount', 'oldbalanceOrg', 'newbalanceOrig',
'oldbalanceDest', 'newbalanceDest',
                    'orig_transactions_count',
'dest_transactions_count']

# Scaling the data
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data[numeric_features])

K = range(1, 15)
inertias = []

for k in K:

    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)

    kmeans.fit(scaled_data)
    #the inertia will measure how well a dataset was clustered by K-
Means
    inertias.append(kmeans.inertia_)

# Plotting the elbow graph
plt.figure(figsize=(16, 8))
plt.plot(K, inertias, 'bx-')
plt.xlabel('Number of clusters, k')
plt.ylabel('Inertia')
plt.title('Elbow Method For Optimal k')
plt.show()
```



The optimal cluster is k=5 which as per the data will be ideal to use the five clusters for peer profiling.

The model fetches data from the Fraud csv which is the training data. For this stage of building and combining the model to score transactions, 100% of the Fraud csv data is used as the training data, the data sent when a customer transacts either via an API or a banking app mimicking the banking application, the data is consumed as a test data where the model uses the training data and received data to score that transaction as either fraudulent or not fraudulent.

The model code:-

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from scipy.spatial.distance import mahalanobis
import numpy as np
import csv
dataset = 'Fraud.csv'

data = pd.read_csv(dataset, delimiter=",", header=0)

# Filter to focus on 'CASH_OUT' and 'TRANSFER' types, crucial in fraud
detection scenarios
data = data[data['type'].isin(['CASH_OUT', 'TRANSFER'])]
```

```
# Calculate initial transaction counts for the dataset
data['orig_transactions_count'] =
data.groupby('nameOrig')['nameOrig'].transform('count')
data['dest_transactions_count'] =
data.groupby('nameDest')['nameDest'].transform('count')

# Define the features to be used for modeling
numeric_features = ['amount', 'oldbalanceOrg', 'newbalanceOrig',
'oldbalanceDest', 'newbalanceDest', 'orig_transactions_count',
'dest_transactions_count']

# Apply scaling to the numeric features
scaler = StandardScaler()
data[numeric_features] = scaler.fit_transform(data[numeric_features])

# Initialize and fit the KMeans model
kmeans = KMeans(n_clusters=5, n_init=10, random_state=42)
data['cluster'] = kmeans.fit_predict(data[numeric_features])

# Compute the centroids and the inverse covariance matrix for Mahalanobis
distance calculation
centroids = kmeans.cluster_centers_
cov_matrix = np.cov(data[numeric_features].T)
inv_cov_matrix = np.linalg.inv(cov_matrix)

# Calculate Mahalanobis distance and flag  fraudulent transactions
data['mahalanobis'] = [mahalanobis(row, centroids[int(cluster)],
inv_cov_matrix) for row, cluster in zip(data[numeric_features].values,
data['cluster'])]
distance_threshold = 2
data['fraudulent_trxn'] = (data['mahalanobis'] >
distance_threshold).astype(int)

# Mapping of 'nameOrig' to clusters for reusing in incoming entries
nameOrig_cluster_map = data[['nameOrig',
'cluster']].drop_duplicates().set_index('nameOrig')['cluster'].to_dict()

# Function to process an incoming entry, dynamically calculating transaction
counts
def process_incoming_entry(entry, data, scaler, kmeans, centroids,
inv_cov_matrix, numeric_features, distance_threshold):
    orig_count = data[data['nameOrig'] == entry['nameOrig'].iloc[0]].shape[0]
+ 1
    dest_count = data[data['nameDest'] == entry['nameDest'].iloc[0]].shape[0]
+ 1
    entry['orig_transactions_count'] = [orig_count]
```

```
    entry['dest_transactions_count'] = [dest_count]

    entry_scaled = entry.copy()
    entry_scaled[numeric_features] =
scaler.transform(entry_scaled[numeric_features])

    if entry['nameOrig'].iloc[0] in nameOrig_cluster_map:
        cluster = nameOrig_cluster_map[entry['nameOrig'].iloc[0]]
    else:
        cluster = kmeans.predict(entry_scaled[numeric_features])[0]

    mahalanobis_distance = mahalanobis(entry_scaled[numeric_features].iloc[0],
centroids[cluster], inv_cov_matrix)
    is_fraudulent = True if (mahalanobis_distance > distance_threshold) else
False

    return {'cluster': int(cluster), 'mahalanobis_distance':
float(mahalanobis_distance), 'fraudulent_trxn': bool(is_fraudulent)}

def validate_transaction(transaction):

    incoming_entry = pd.DataFrame([transaction])
    # Process the incoming entry
    result = process_incoming_entry(incoming_entry, data, scaler, kmeans,
centroids, inv_cov_matrix, numeric_features, distance_threshold)

    return result
```

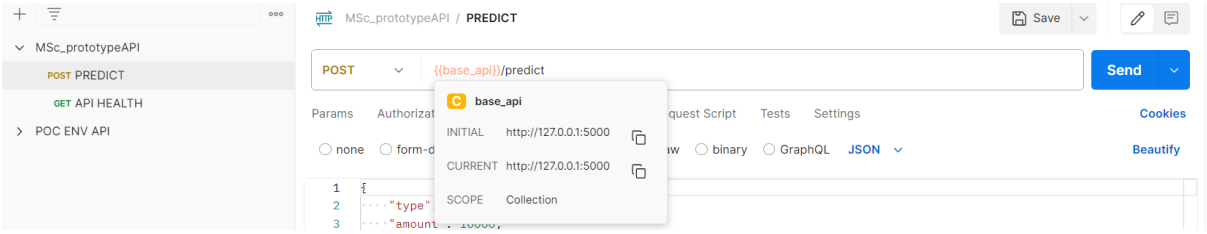
To test the model and confirm it is working well, used an API which mimicks a banking app to send transactions and confirm the model validity.

API Code:-

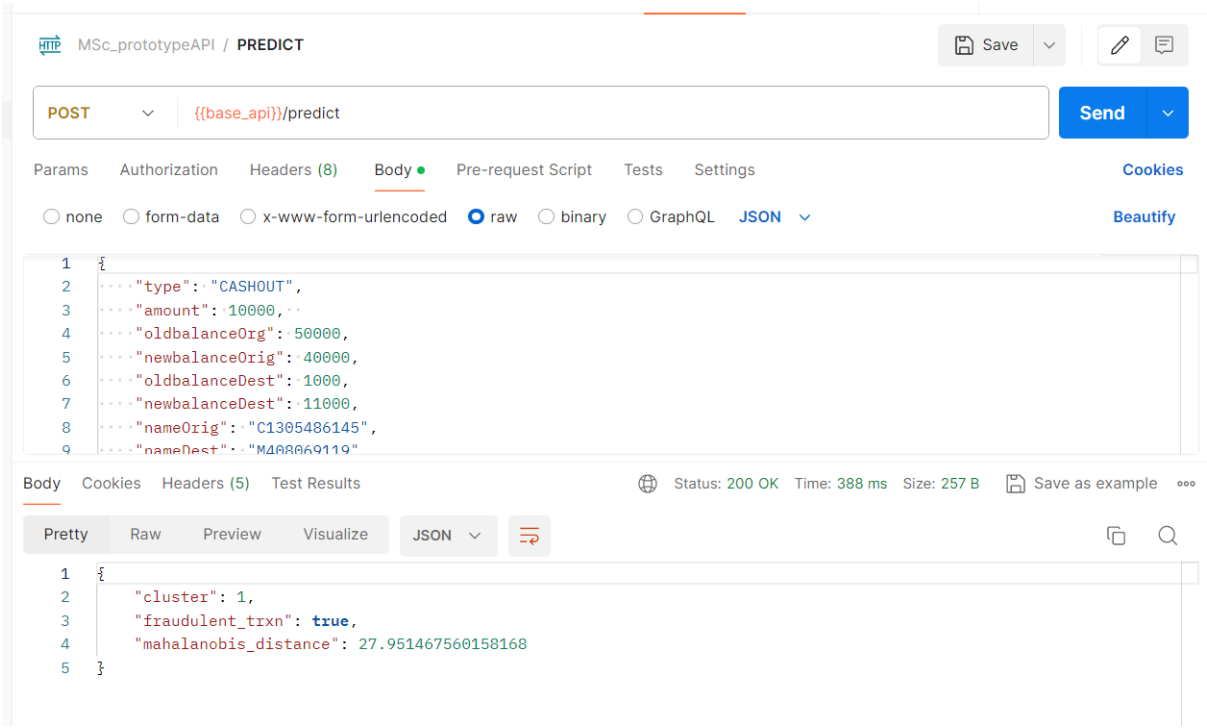
```
from flask import Flask, request, jsonify,render_template
import pandas as pd
import mysql
import mysql.connector
import json
import model_prototype

# Load the pre-trained model

app = Flask(__name__,template_folder='templates')

@app.route("/health")
```

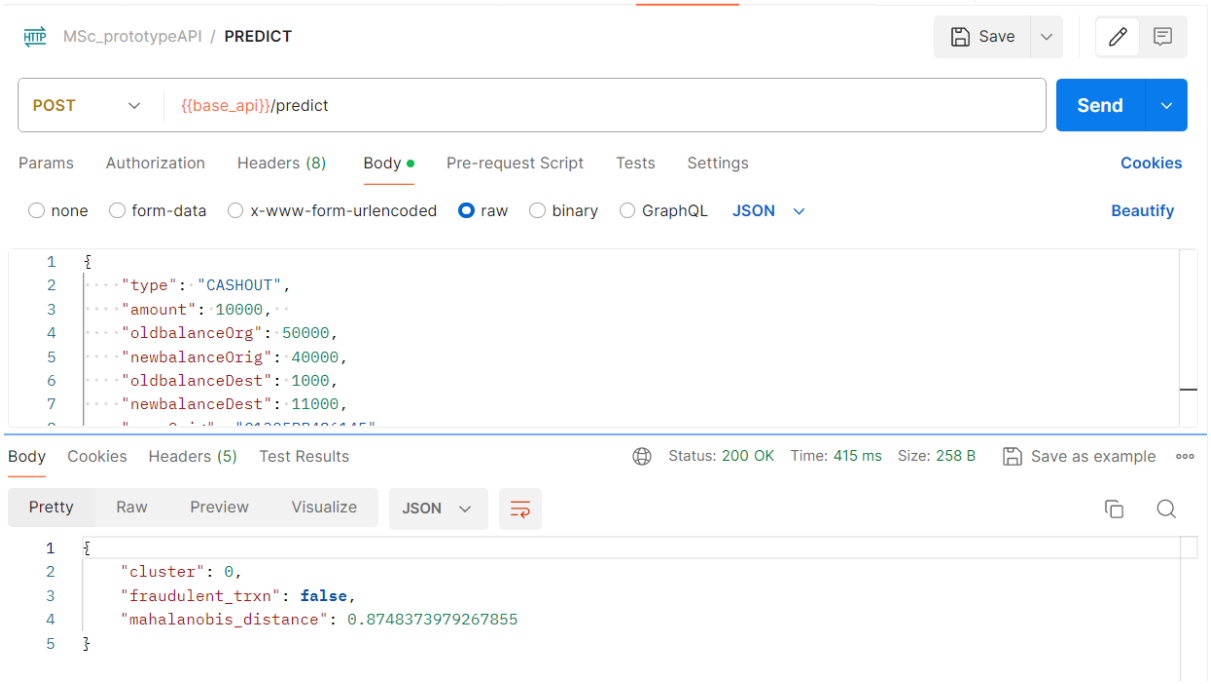
```
def health_check():
    return jsonify({"message": "API is healthy!"})

@app.route('/transactions')
def view_transactions():
    # Connect to your database
    conn = mysql.connector.connect(
        host="localhost", user="root", password="root", db="larry"
    )
    cursor = conn.cursor()
    # Query Transactions
    cursor.execute("SELECT id, data, is_fradulent,amount FROM transactions")
    transactions = cursor.fetchall()
    # Close database connection
    conn.close()
    return render_template('transactions.html', transactions=transactions)

@app.route("/predict", methods=["POST"])
def predict_fraud():
    # Get transaction data from request
    data = request.get_json()

    response = model_prototype.validate_transaction(data)

    conn = mysql.connector.connect(
        host="localhost", user="root", password="root", db="larry"
    )
    cursor = conn.cursor()
    data_to_save = {
        'amount': data['amount'],
        'data': data,
        'is_fradulent':response['fraudulent_trxn']
    }

    json_data = json.dumps(data_to_save)
    # Insert data into database
    is_fraudulent_int = 1 if response['fraudulent_trxn'] else 0
    cursor.execute("INSERT INTO `transactions` (`data`,`amount`,
`is_fradulent`,`create_stamp`) VALUES
('"+json_data+"','"+str(data['amount'])+"','"+str(is_fraudulent_int)+"',now())
")
    conn.commit()

    return jsonify(response)

if __name__ == "__main__":
    app.run(debug=True)
```

Postman API:-

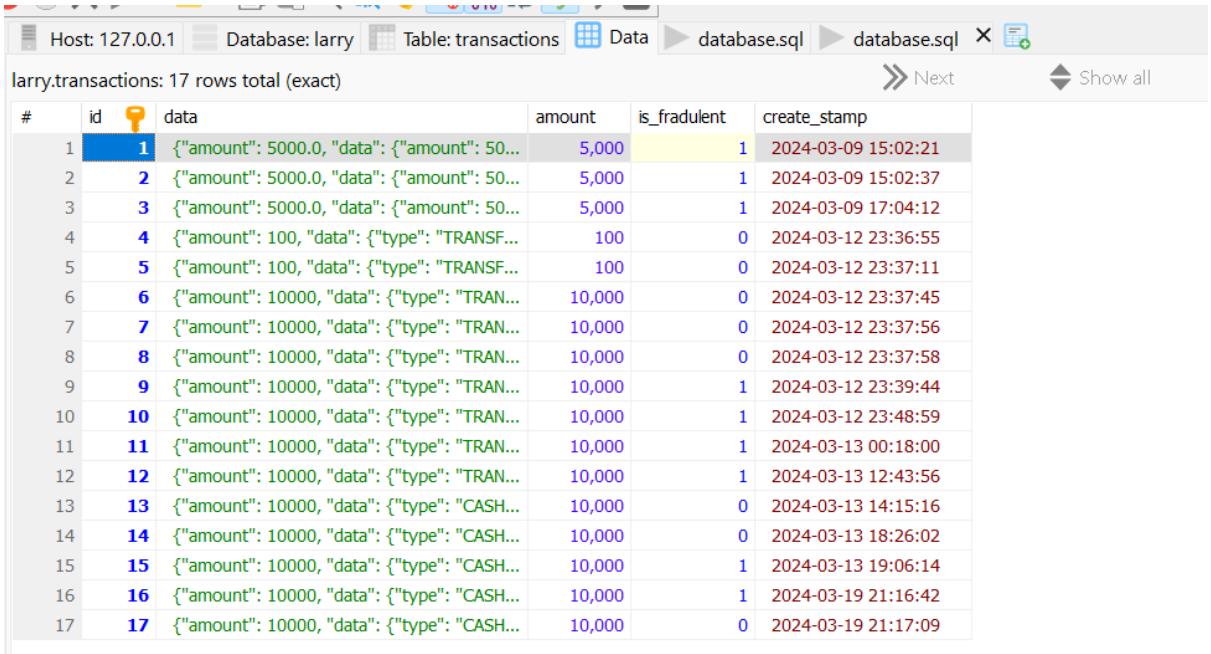


A test of the API to show a scoring test of the model:-





The **MySQL storage** for the scored transactions:-

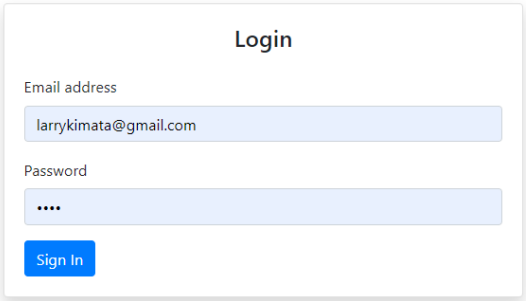


#	id	data	amount	is_fradulent	create_stamp
1	1	{"amount": 5000.0, "data": {"amount": 50...	5,000	1	2024-03-09 15:02:21
2	2	{"amount": 5000.0, "data": {"amount": 50...	5,000	1	2024-03-09 15:02:37
3	3	{"amount": 5000.0, "data": {"amount": 50...	5,000	1	2024-03-09 17:04:12
4	4	{"amount": 100, "data": {"type": "TRANSF...	100	0	2024-03-12 23:36:55
5	5	{"amount": 100, "data": {"type": "TRANSF...	100	0	2024-03-12 23:37:11
6	6	{"amount": 10000, "data": {"type": "TRAN...	10,000	0	2024-03-12 23:37:45
7	7	{"amount": 10000, "data": {"type": "TRAN...	10,000	0	2024-03-12 23:37:56
8	8	{"amount": 10000, "data": {"type": "TRAN...	10,000	0	2024-03-12 23:37:58
9	9	{"amount": 10000, "data": {"type": "TRAN...	10,000	1	2024-03-12 23:39:44
10	10	{"amount": 10000, "data": {"type": "TRAN...	10,000	1	2024-03-12 23:48:59
11	11	{"amount": 10000, "data": {"type": "TRAN...	10,000	1	2024-03-13 00:18:00
12	12	{"amount": 10000, "data": {"type": "TRAN...	10,000	1	2024-03-13 12:43:56
13	13	{"amount": 10000, "data": {"type": "CASH...	10,000	0	2024-03-13 14:15:16
14	14	{"amount": 10000, "data": {"type": "CASH...	10,000	0	2024-03-13 18:26:02
15	15	{"amount": 10000, "data": {"type": "CASH...	10,000	1	2024-03-13 19:06:14
16	16	{"amount": 10000, "data": {"type": "CASH...	10,000	1	2024-03-19 21:16:42
17	17	{"amount": 10000, "data": {"type": "CASH...	10,000	0	2024-03-19 21:17:09

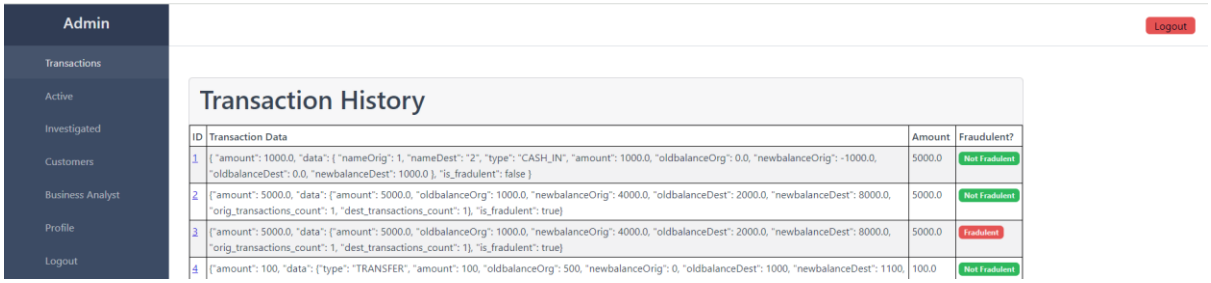
5.2.5 Web Management App Implementation

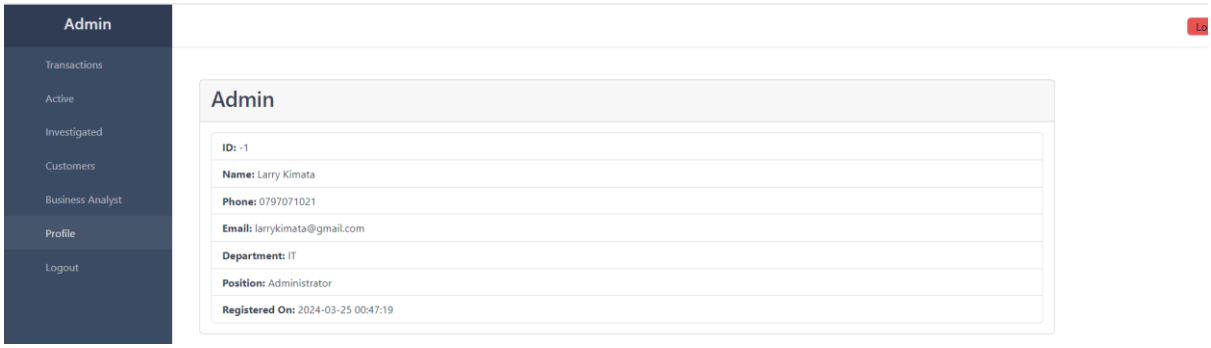
This is the main administration web application that will be used to perform business analyst and customer user creation. The prototype is hosted locally on a laptop to avoid incurring hosting charges that might be expensive.

The Web management application first view looks as below:-



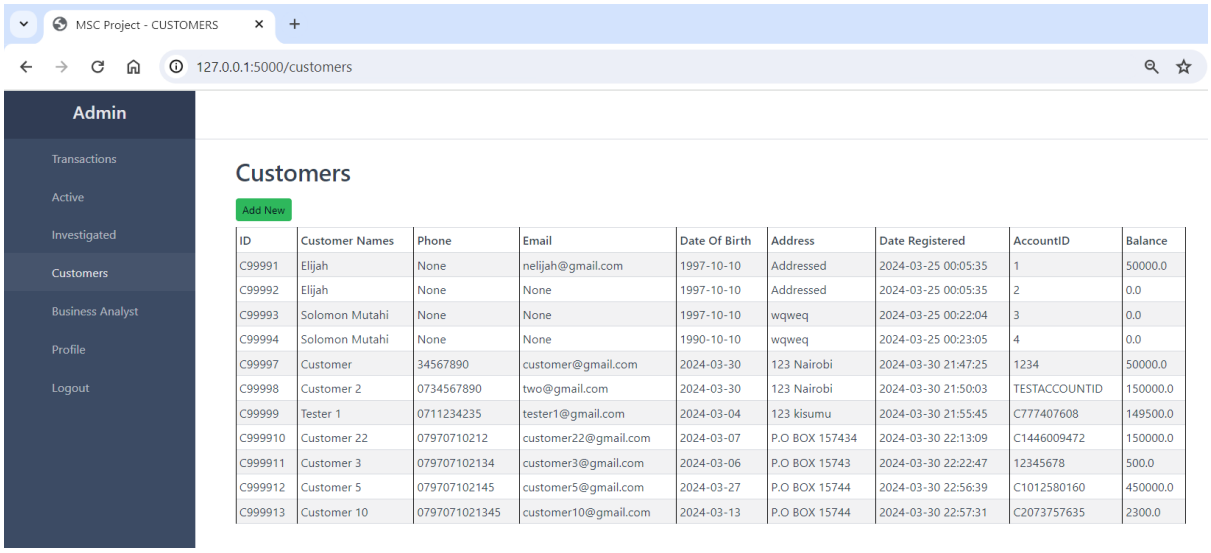
Once login; the administrator will have the below views to now proceed and perform user creations:-

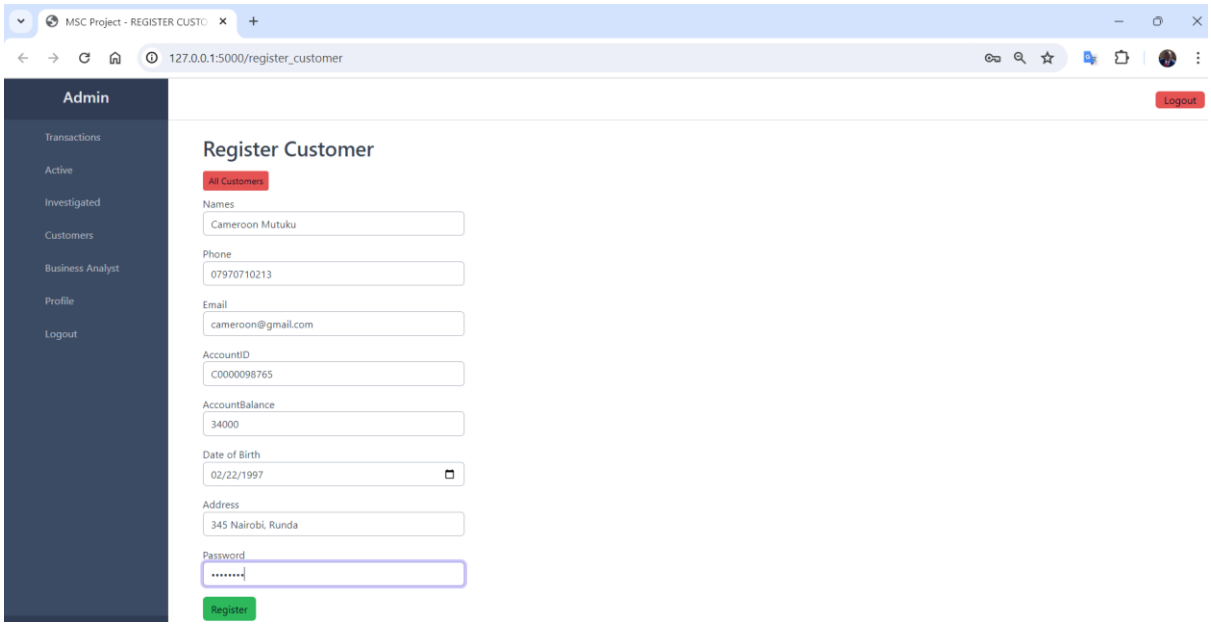




5.2.5.1 Customer Admin View

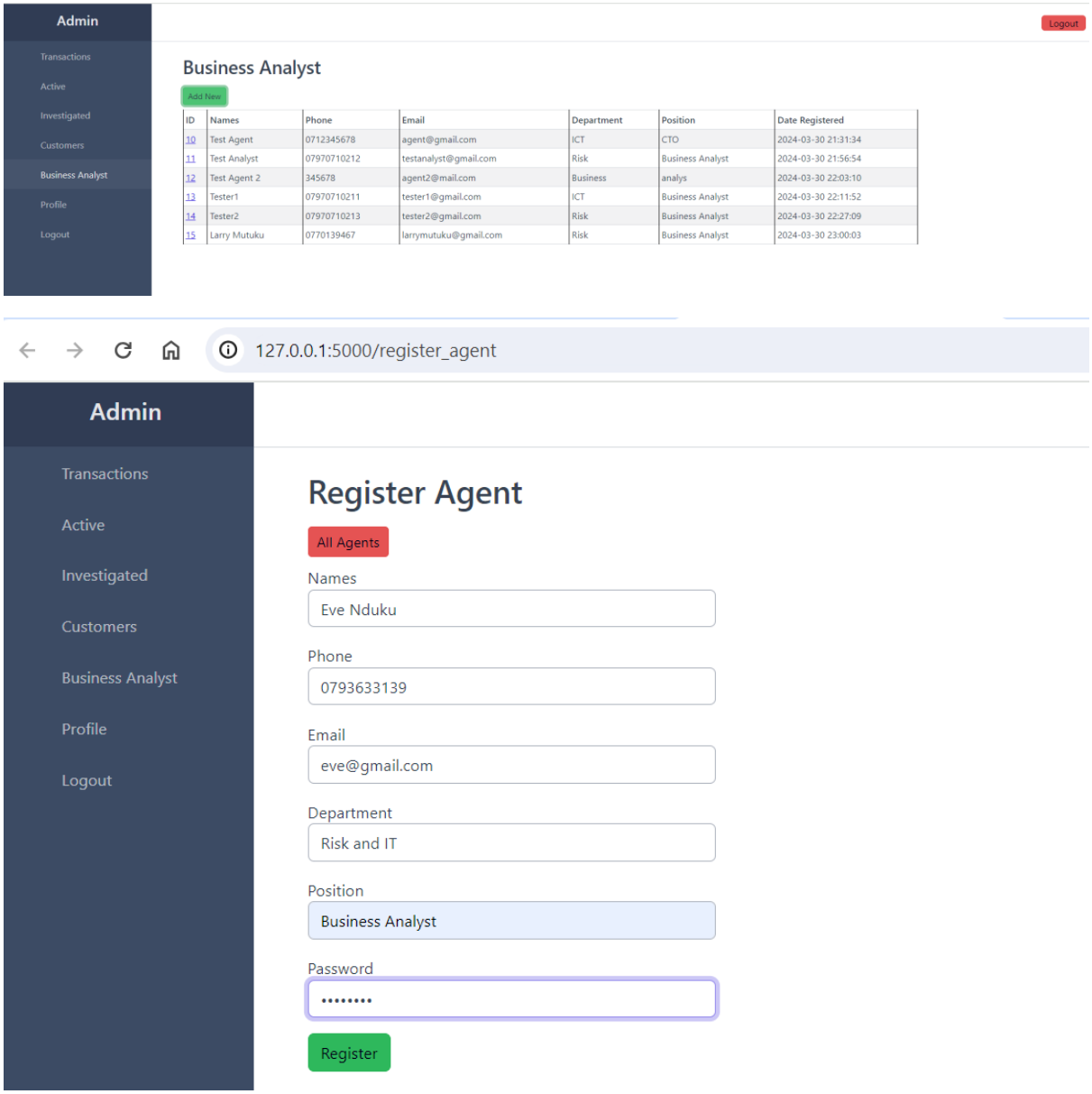
To create a customer, the administrator clicks on customer's where one is able to see all existing customers and an option to add other customers successfully.





5.2.5.2 Business Analyst Admin View

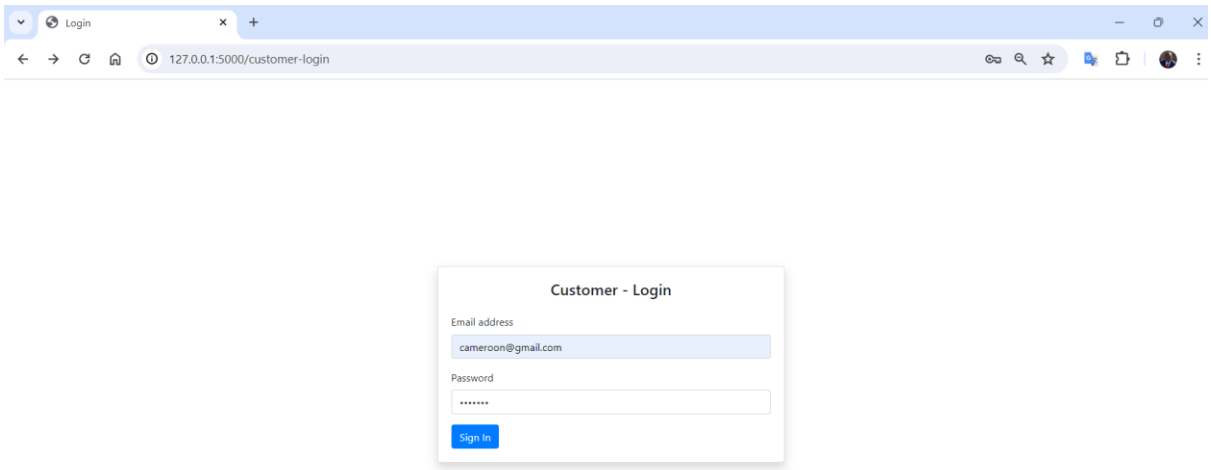
To create a business analyst, the administrator clicks on Business Analyst View where one is able to see all existing business analyst and an option to add other business analysts successfully.



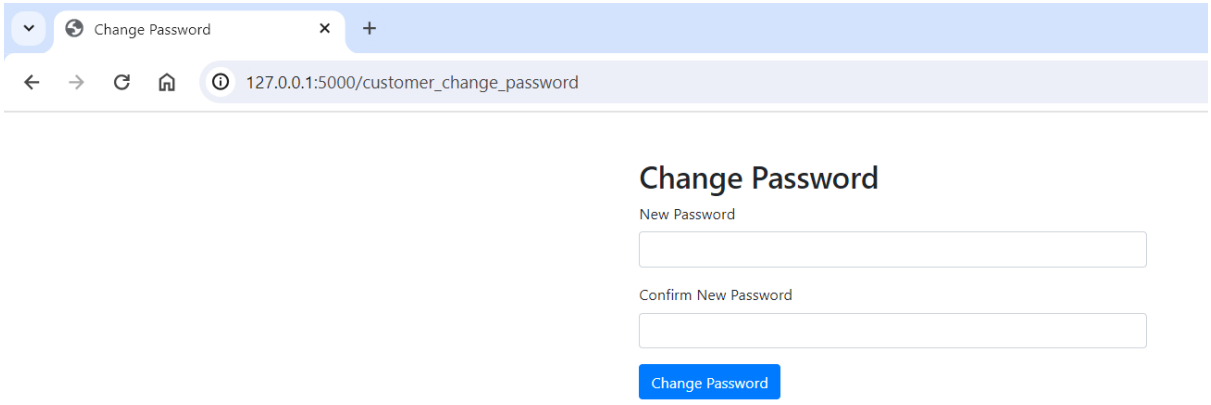
5.2.6 Banking App Implementation

For the Banking App, we have employed use of html and bootstrap framework to create the forms and views that will be used in the internet banking application. The other code the interfaces between the html and bootstrap to make the API calls by the banking application to be available is used of python code assisted with other machine learning libraries like sklearn, pandas, SciPy, flask and use of Json.

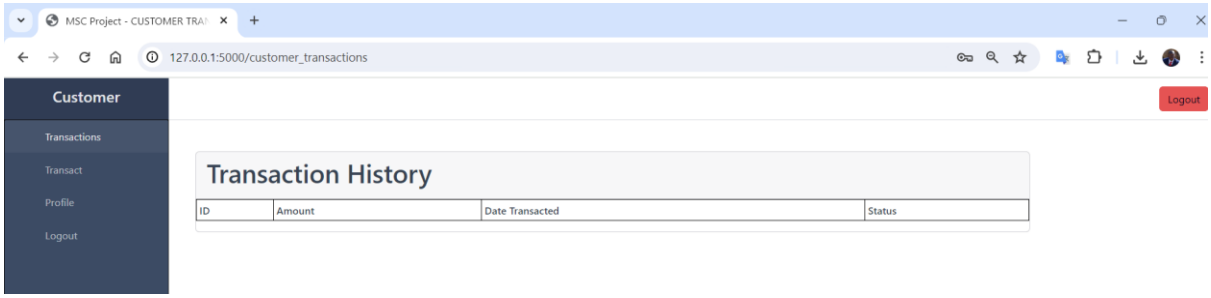
After the Login details of the customer have been created, the customer uses this login to access the banking application.



As a security precaution, all the customers accounts are forced to change their password after the first login.

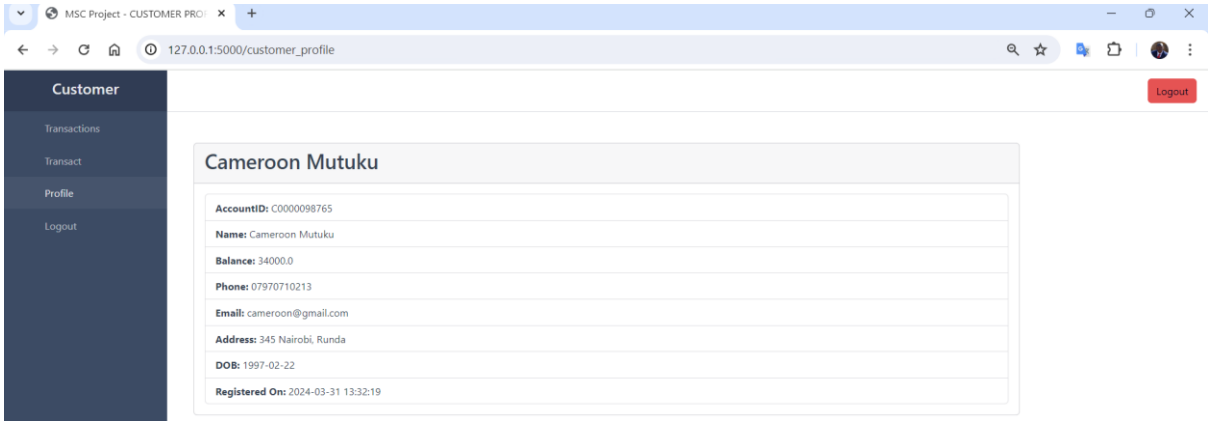


Once they change their password, they can see the banking application.

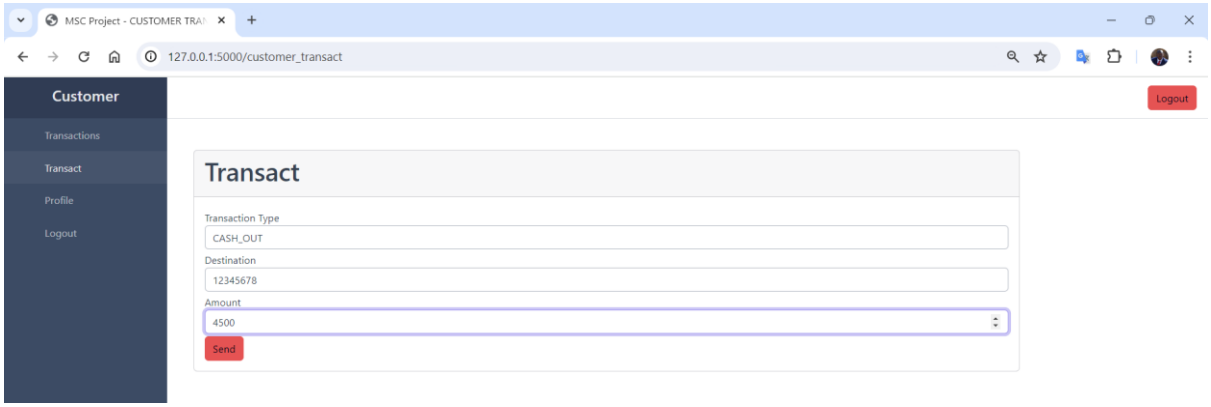


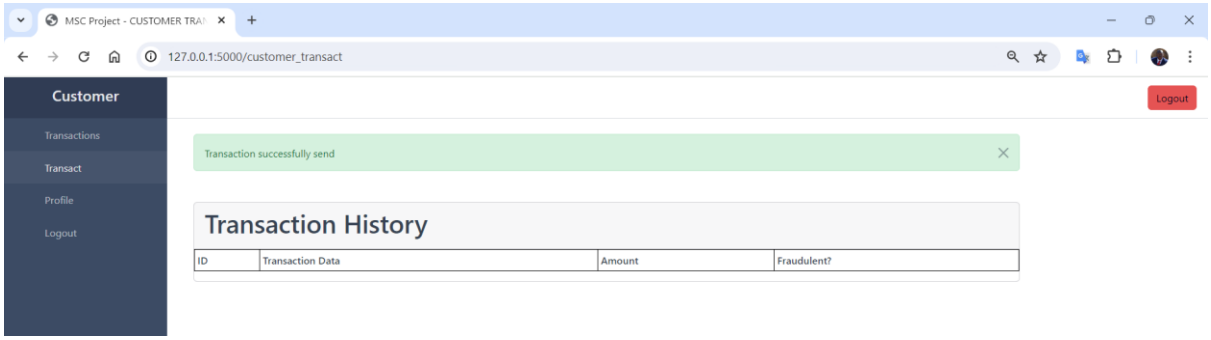
Transactions contains all the transaction that the user has done

The Profile shows the profile of the customer.



The transact view allows the customer to perform transactions.



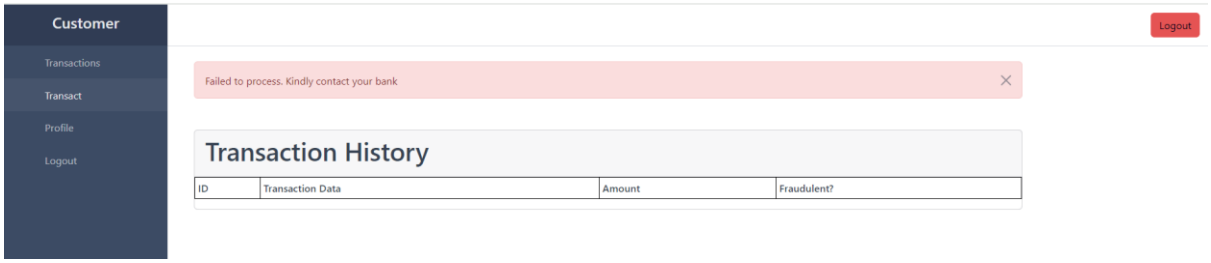


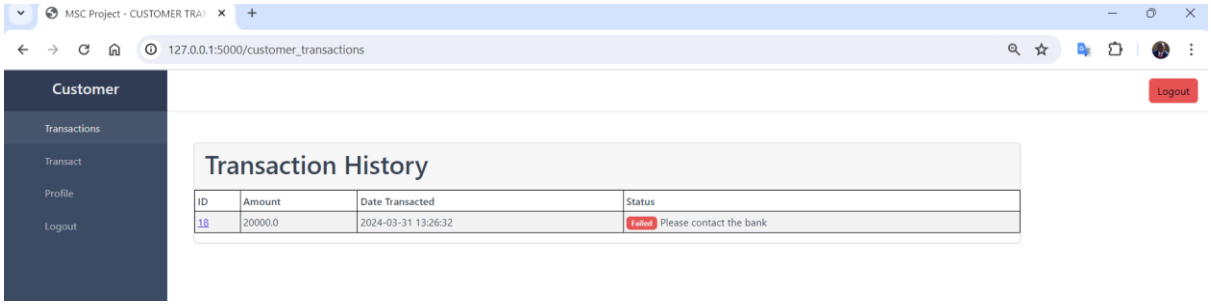
Customer can choose the transaction type, the recipient and the amount.

Once the correct transaction details are chosen , customer can click on the send button for transaction to be send.

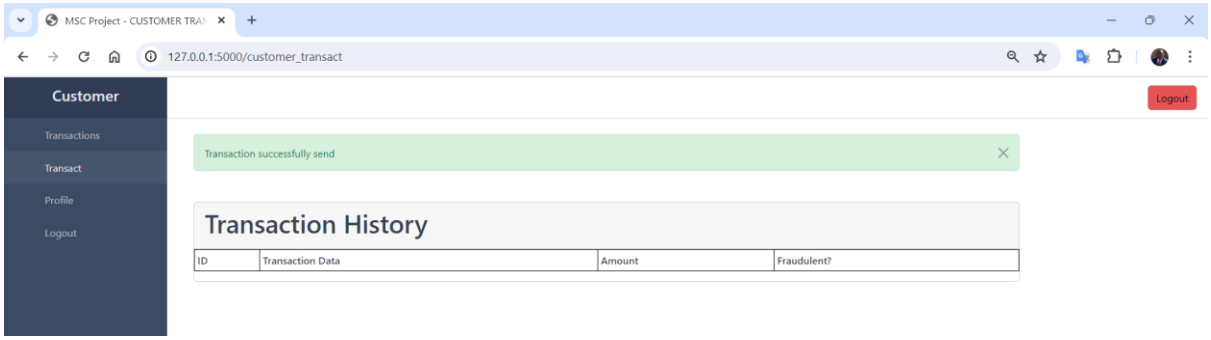
Transaction send to the recipient must first pass through the model, if the model scores the transaction as genuine, the recipient will receive the money, but if the transaction is fraudulent, the model will flag as fraudulent and this will be blocked thus the transaction will not be received by the recipient.

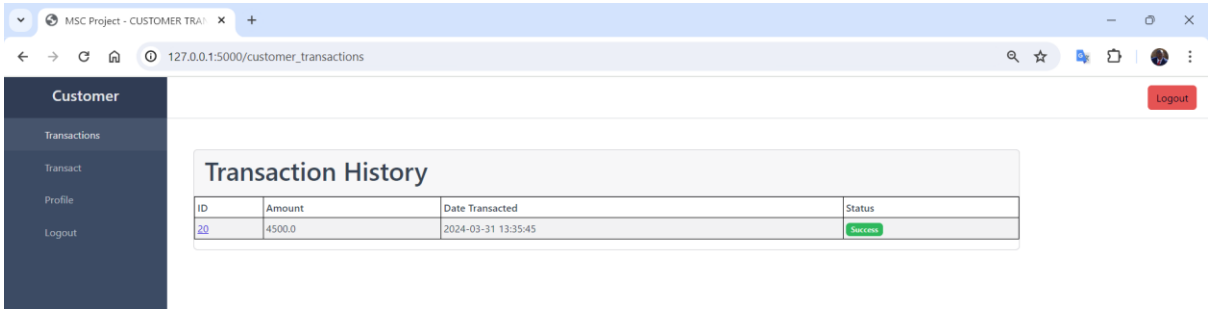
Fraudulent Transaction





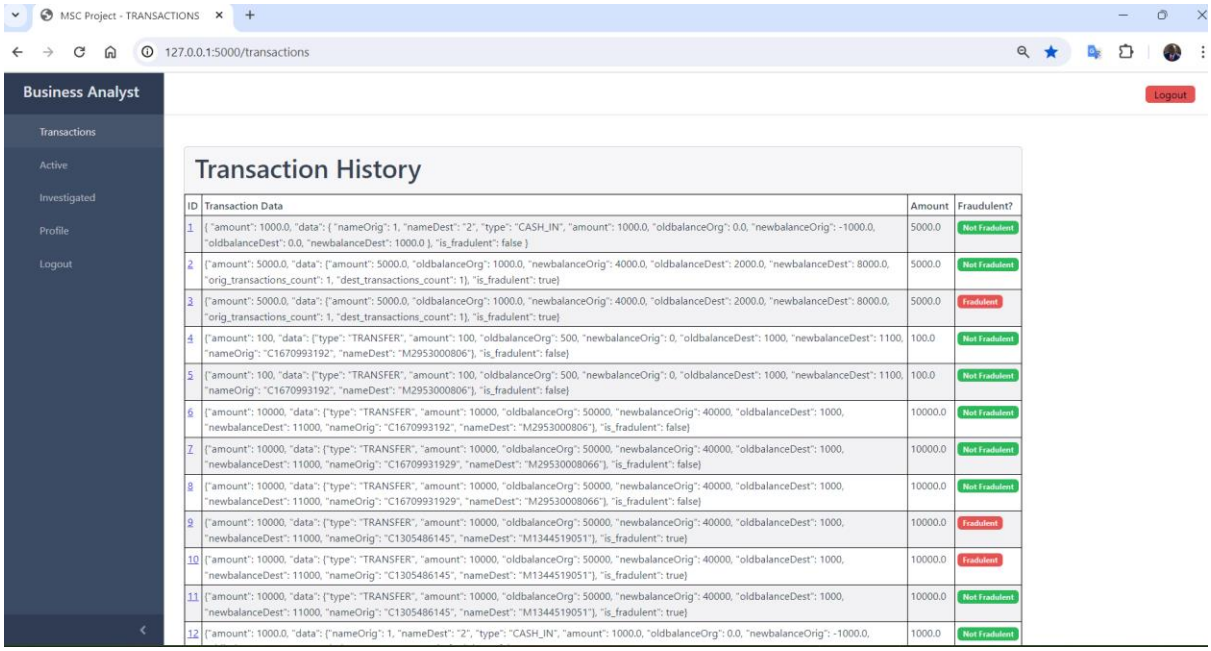
Genuine Transaction





5.2.7 Business Analyst Implementation

The business analyst logins in using the below portal and has the below view:-

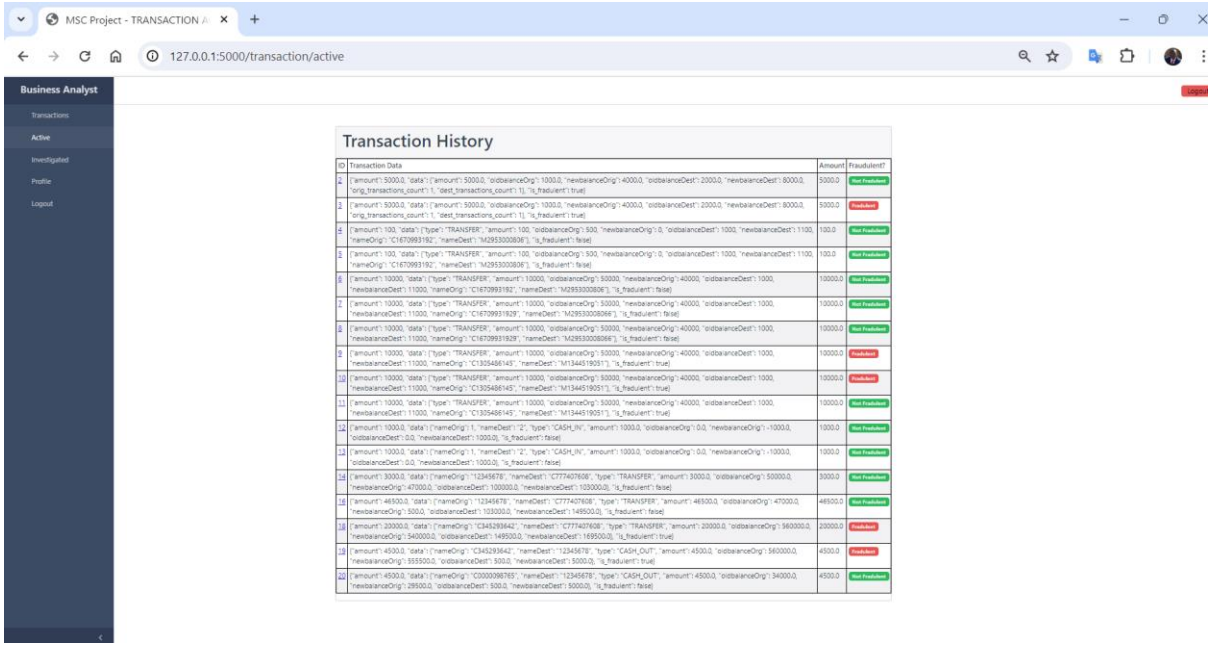


Transactions: Shows all the transactions that have been scored by the model.

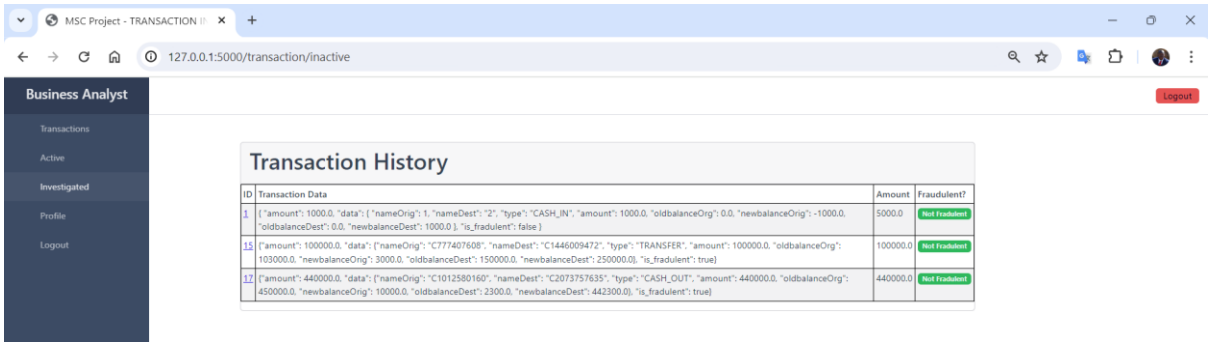
Active: Shows the transactions that have not been scored by the model.

Investigated: Shows the transactions that have been investigated by the business analyst and commented.

Active View



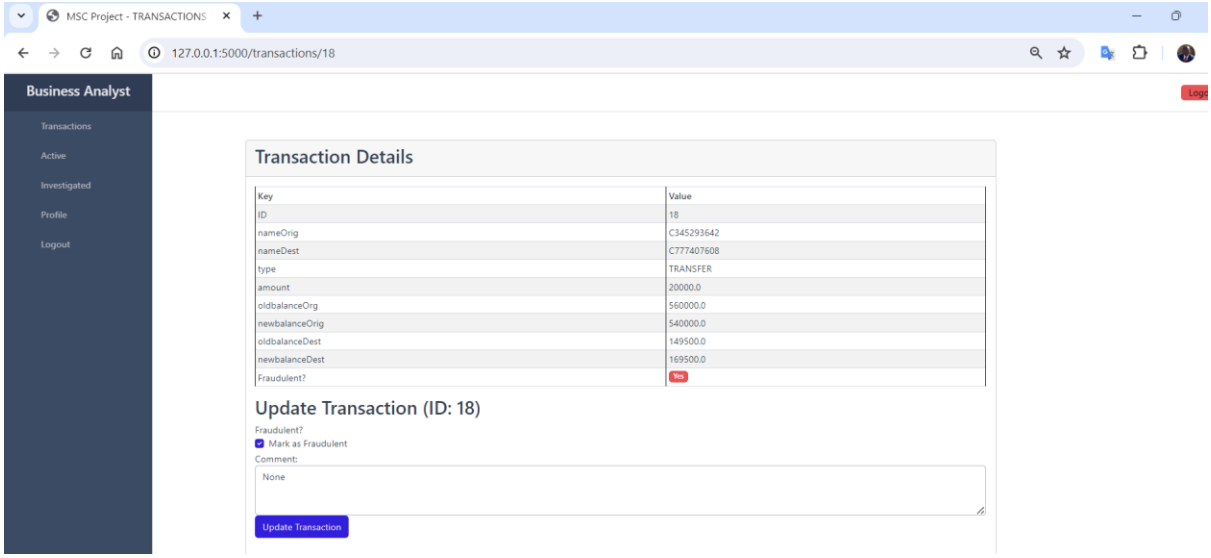
Investigated:



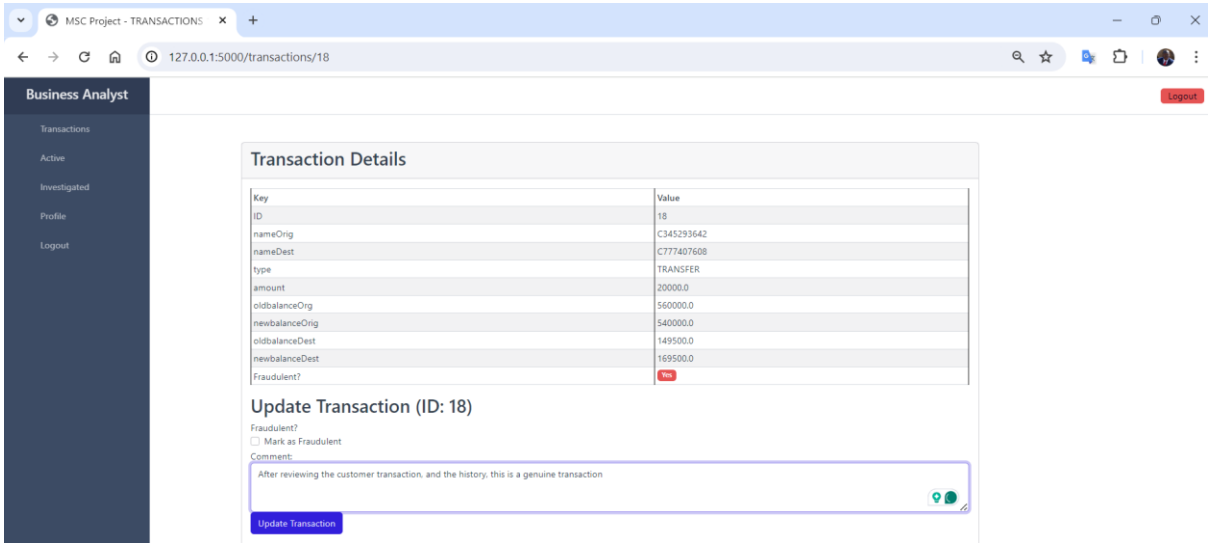
Business Analyst Workflow

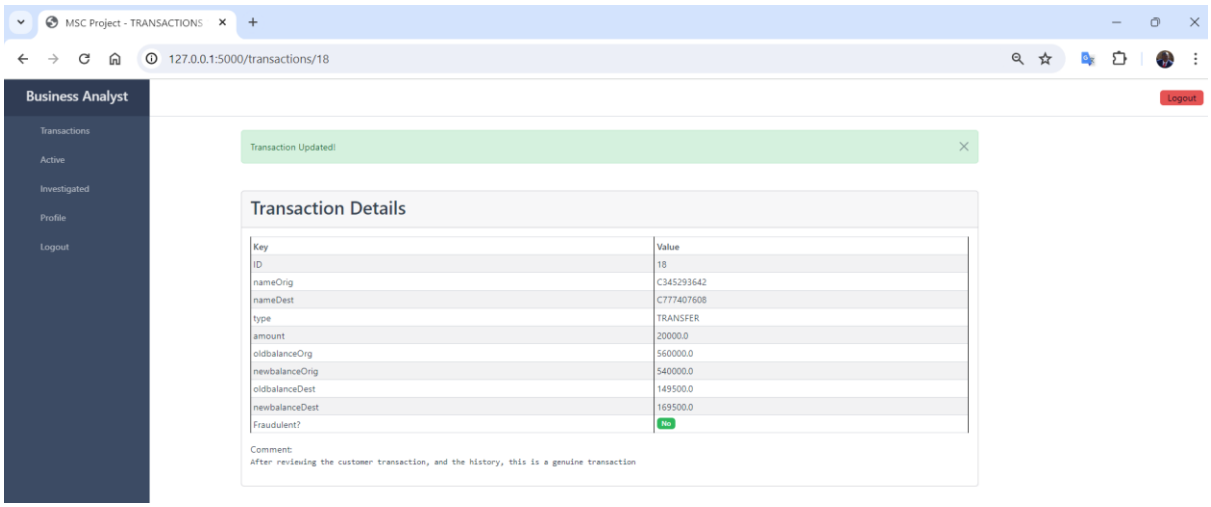
A business analyst will click on the ID of the transaction, which will open the transaction details:-

An example is ID 18 which is a transaction that was blocked:-



A business analyst has reviewed the transaction and reviewed the customer's historical transactions and concluded that the transaction is not fraudulent, they will unclick on the mark as fraudulent , put a comment and update the transaction.

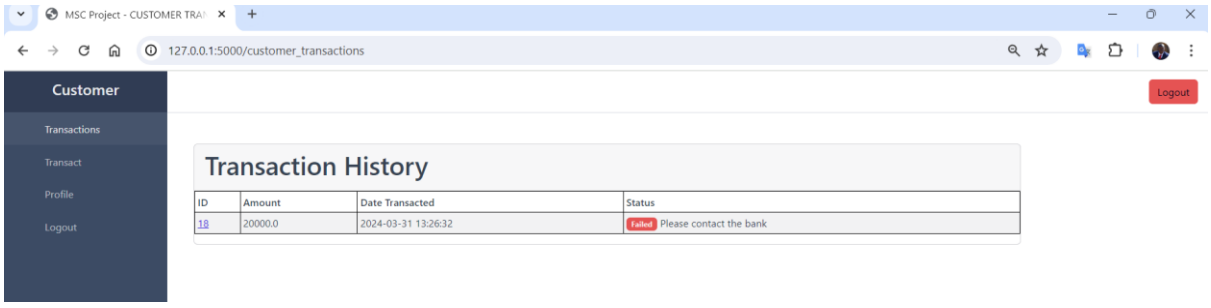




By clicking on the update, the transaction has been unblocked and sent to the recipient successfully.

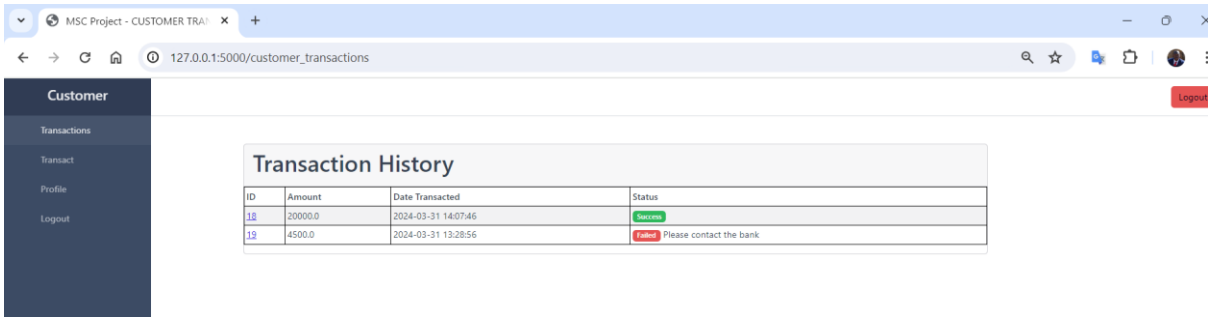
On the customer's profile who the transaction was blocked is as below:-

Before, when transaction was blocked:



After, when the business analyst has reviewed and concluded the transaction is genuine.

Transaction ID 18 has been processed and committed.



5.3 System Testing

This section outlines the tests that were performed on the application prototype to validate their functionality, reliability, and performance. The tests used are validation, functionality, and usability tests.

5.3.1 Functional Testing

Table 5 Functional Testing: Banking App User Transacts

Test Case Name: Banking App user transacts		Test Case: 1	
Description: The user transacts on the banking app which is an internet banking of the application			
Pre-Condition: The user should have visited the application			
Step	Action	Expected Results	Pass/Fail
1	user opens the banking app	The application loads the sign in/login page	Pass
2	User can login	Application can validate only users in the database	Pass
3	User can transact by sending money	Application allows fetch and input of user details for a transaction to be committed	Pass
4	Transaction is successful	Transaction is sent from banking app to the model using API and stored in the Database	Pass
Post-Condition: User can login and transact			

Table 6 Functional Testing: Web Application User Access

Test Case Name: Web Application user Access		Test Case: 2	
Description: The admin or business analyst can login to the web application and interact			
Pre-Condition: The user should have visited the application			
Step	Action	Expected Results	Pass/Fail
1	user opens the web application	The application loads the sign in/login page	Pass
2	User can login	Application can validate only users in the database	Pass
3	User can create users for the mobile app which is a mobile bank prototype and the web app	Application can create users successfully	Pass
Post-Condition: User can access both web application and interact easily			

5.3.2 Validation Testing

Through the validation of users, for this testing , the prototype achieved a 95% validation testing.

Table 7 Validation Testing

Test Case Name: Validation of the application- Prototype Application		Test Case: 3	
Description: Users interacting with the application to check if its user friendly and UI is intuitive			
Pre-Condition: The user should have visited the application and used it			
Step	Action	Expected Results	Pass/Fail
1	System accuracy and speed	The application loads all the pages and interaction. The user input is validated and ensure correct output and takes less time	Pass
Post-Condition: User can access the application prototype and interact easily			

5.3.3 Usability Testing

Usability Testing was done to confirm the user friendliness of both the prototype application.

Table 8 Usability Testing

Test Case Name: Usability of the Application		Test Case: 4	
Description: Test to gauge the overall usability of the application			
Pre-Condition: The user should have visited the application and used it			
Step	Action	Expected Results	Pass/Fail
1	User can access the user interface, the web application menu items and interact with them easily	Prototype application interfaces should be clear and concise	Pass
2	User can navigate the application with ease	Menu items should be intuitive, clickable, responsive and descriptive	Pass

5.3.4 Compatibility Testing

The compatibility testing was done to make sure that the web application works well with existing web platforms.

Table 9 Compatibility Testing

Web Browser	**Compatibility**
Microsoft Edge	Yes
Mozilla Browser	Yes
Google Chrome	Yes
Safari	Yes

6 Chapter6: Discussions and Key findings

6.1 Introduction

As fraud becomes more sophisticated and harder to detect, swindlers and fraudsters are active and this way, financial institutions are increasingly finding it harder to combat these fraudulent tactics that lead to loss of money.

The research study of this project was aimed at coming up with a solution that can detect fraudulent transactions by checking the distinctive attributes of peers and using it to check if a fraud is genuine or fraud. The study uses peer profiling and application of mahalanobis as a distance measure to detect frauds.

To develop the prototype, the research study sought to comprehend the challenges which are faced by financial institutions in relation to fraud so that this research could incorporate that when building a good model and a system that would assist on detection of the fraudulent transactions. This was achieved through using interviews where the respondents provided the responses which were analyzed and visually represented.

6.2 Primary Discoveries and Achievements

The literature review used while performing this study shows that digital banking fraud is still an issue that is affecting financial institutions. The different ways of detection of these fraudulent transactions can range from finding a pattern, a machine learning model, or basic rules. This research study aims at using a peer profiling model that employs use of mahalanobis distance as the multivariate distance analysis to detect fraud.

A key finding is that it is necessary to find the optimal number of clusters before performing peer profiling. As development of the prototype was ongoing, this was key to make sure that model gets the best accuracy. The use of elbow method which is a graphical method for finding the optimal value of K as we are using K-means clustering algorithm was key was used to get the best value of K for great performance of the model.

The elbow method on the training data showed visually that K=5 was the optimal number thus we had to use 5 clusters.

The second key discovery was that it was possible to score transactions via the model in less than three seconds while the model loaded all the training data and scaling the numerical features which were used to get the behavior attributes of the customer and score the transaction. This is very fast as the training data contains more than a million transactions.

The third key discovery is that developing the model to map where the customer is contained in which cluster was key so that whenever a transaction was sent to the model, it would get where the customer is mapped to which cluster and the centroid of the cluster used and where the transaction falls to get the mahalanobis distance. This was an achievement as in every model run, we do not need to place the existing customer in a cluster again, rather get the cluster position and use that to score the transaction. The below code was able to achieve that and reduce time in seconds which was used to score the transaction.

```
if entry['nameOrig'].iloc[0] in nameOrig_cluster_map:
    cluster = nameOrig_cluster_map[entry['nameOrig'].iloc[0]]
else:
    cluster = kmeans.predict(entry_scaled[numeric_features])[0]
```

In this context and research study, with use of a multivariate distance to score transactions, Mahalanobis distance showed a higher accuracy compared to Euclidean distance as well as Manhattan distance which is a key discovery and achievement.

The development of a model that can segment customers into peer groups and clusters based on these attributes which are the transaction amount, the balances, transactions count of both sender and receiver was a key achievement to show how this is possible. This led to achievement of a higher accuracy compared to other detection methods of fraudulent transactions.

6.3 Exploration of the Research Objectives

The first objective has been achieved; the model has incorporated the use of peer profiling by grouping the transactions on a unique identifier of nameOrig and used k-means to cluster the individuals of the transactions to the clusters depending on their transaction attributes and behavior. K-means assisted to group them in 5 clusters which was optimally achieved by using elbow method to find the best number of clusters to be used. Mahalanobis multivariate analysis was used to calculate the distance between the centroid where the cluster is and where the transaction placed by the model thus able to detect as fraudulent or a genuine transaction.

The model accuracy was achieved by comparing the data's fraud field and what the model scores is 92% which shows a high accuracy percentage of the model, compared to the works that have been done; thus, showing that the system is able to achieve this objective.

The second objective which an interview was used to investigate the challenges faced by financial institutions due to fraudulent activities so that this would help more in finding how we can further improve the model to get these fraudulent transactions. This was done and assisted to improve the model by understanding the challenges faced. From the interview used, it was concluded that account takeovers is the highest form of fraud used in financial institutions were one is able to access the account and perform transactions. By using this prototype, it will be easy to catch fraud and block as incase one takes an account of someone

they will not able to perform pattern like transactions as one does thus the transaction will be detected as fraudulent which makes the prototype to be ideal for many financial institutions.

The third objective which was to review the current used methods in mobile money fraud detection was done in the second chapter of the literature review thus this was achieved. It was seen that the current methods employ use of simple patterns and other distance measures which have low accuracy. The application of mahalanobis distance in the peer profiling shows that the MD distance is ideal and has the highest accuracy compared to other distances.

The fourth objective which entails training the model by developing and use of mahalanobis distance. This has been achieved by seen the below code, which entails the model been trained on the data to be ready to score transactions and further using mahalanobis distance measure as the metric to flag transactions as either fraudulent or genuine.

The below segment of code shows the numerical features of the train data to be used in model score and later application of mahalanobis distance to be used as the multivariate distance analysis; thus, this objective has been met.

```
# Define the features to be used for modeling
numeric_features = ['amount', 'oldbalanceOrg', 'newbalanceOrig',
'oldbalanceDest', 'newbalanceDest', 'orig_transactions_count',
'dest_transactions_count']

# Apply scaling to the numeric features
scaler = StandardScaler()
data[numeric_features] = scaler.fit_transform(data[numeric_features])

# Initialize and fit the KMeans model
kmeans = KMeans(n_clusters=5, n_init=10, random_state=42)
data['cluster'] = kmeans.fit_predict(data[numeric_features])

# Compute the centroids and the inverse covariance matrix for Mahalanobis
distance calculation
centroids = kmeans.cluster_centers_
cov_matrix = np.cov(data[numeric_features].T)
inv_cov_matrix = np.linalg.inv(cov_matrix)

# Calculate Mahalanobis distance and flag  fraudulent transactions
data['mahalanobis'] = [mahalanobis(row, centroids[int(cluster)],
inv_cov_matrix) for row, cluster in zip(data[numeric_features].values,
data['cluster'])]
```

The last objective of this research study entailed to validate the model by using mobile money transaction dataset and comparing the fraud label on the data.

```
from sklearn.metrics import accuracy_score

# Calculate metrics
test_accuracy_percentage = accuracy_score(test_data['isFraud'],
test_data['fraudulent_trxn']) * 100

# Print metrics
print("\nTesting Data Metrics (as percentages):")
print(f"Accuracy: {test_accuracy_percentage:.2f}%")
```

On the above code, isFraud is the original label for the received dataset. The above code used will compare what the original data label is compared to how the model-built scores the transactions. By using this, below is the accuracy:-

```
Testing Data Metrics (as percentages):
Accuracy: 92.36%
```

With the above code and further explanations, this meets the fourth objective under the specific objectives thus the model has been validated to score accurately as to what is within the data label isFraud; thus, achieving the above-mentioned high accuracy.

6.4 Evaluation of the Prototype

The prototype that has been developed has number of pros as compared to other existing systems. The prototype develops the use of simple to understand logic which can be comprehended easily in the financial institutions and has showed how to be able to detect fraudulent transactions.

Below are the pros and cons for this prototype.

Advantages of the prototype

i. Prototype can train on historical patterns of the customer and use them to score when transactions are sent to the model.
ii. The prototype can successfully allow a transaction to be committed if its genuine and block if its fraudulent.
iii. If it's a false positive and detected as fraudulent while its genuine, the prototype allows the business analyst of the financial institution to review the transaction and if so, mark as genuine and the transaction will be automatically released and committed.
iv. The mobile app which mimics a bank's banking app can communicate with the model and data is stored in a database, MySQL.
v. Business analyst has overview and defined workflows on how to investigate a case and deem it as genuine or fraud leading to another action.

vi. The prototype achieves a research study that has not been done; and that is peer profiling with the use of mahalanobis distance which is a multivariate distance analysis. With this study, the achievement and finding are that the model has a higher accuracy compared to other systems having a lower accuracy.

Disadvantages of the prototype

For the very big financial institutions which work with a lot of Big Data and need scoring to be done using milliseconds , addition of Apache Spark would be ideal to make the prototype more with powerful processing capabilities and faster return of response. This means more resources will be needed for the environment that this will be hosted, i.e more RAM, CPUs, and a high availability storage; which means this could be expensive to acquire the environment; this is an area that can be focused on as a future advancement for this research study and prototype.

7 Chapter7: Recommendations, Future Work and Conclusions

7.1 Recommendations

Ingestion of the data: For the prototype to work effectively, the organization using this prototype especially financial institutions would need to ingest their data to the model so that the model can be trained; to assist score the transactions better.

Optimal Clusters: The model has shown a very good accuracy by using elbow method to find the optimal number of clusters. This also would be an ideal step to be done to get the number of clusters to be used when peer profiling the transactions.

7.2 Future Work

For future work, companies using Big Data need a higher computing and processing capacity, where this model would be slower to their day-to-day operations.

Below are the key additions that can be done for future work:-

Algorithm Improvements:- A concept revolving use of Elastic search to store the data, where indexing operations are done easily and use of Apache Spark to effectively run the model and score the transactions would be the best ideal for the big companies. This is research that can be done on top of what this model does to achieve a higher processing rate, a lower response time to financial institutions which have high transactions per second and use Big Data.

Application in Diverse Domains:- The prototype can also be used in different fraud setups other than in financial institutions to even procurement frauds or healthcare frauds which can assist in combating those types of frauds.

Retraining Model with transactions marked by business analysts: As the business analysts can investigate the transact and deem if genuine or fraudulent which is correct. There is a need for a script to be developed that will send this data to the fraud csv every day so that the model can be retrained to achieve a higher accuracy when scoring transactions.

7.3 Conclusions

Effectiveness of peer profiling: By using peer profiling, the model was able to accurately group customers to different clusters accurately, hence an accuracy of 92%. This accuracy was comparing what the original dataset had on isFraud field to what the model scored thus this was an effective way.

Effectiveness of the Mahalanobis distance as the measurement criteria: The distance measure used was able to account for the covariance among variables which offered a better way to calculate distance between the centroids to the transaction's cluster position thus having a better accuracy that the other compared distances which had 82%, 79% for Euclidean distance and Manhattan distance respectively.

Impact on the field: Fraud is a very key issue where financial institutions lose a lot of money as it's a loophole where many have not found a way to patch it. This prototype can be able to bridge the gap between and provide a solution that will be able to detect and prevent transactions as fraudulent or not fraudulent so that the bank can make an informed decision.

References

Allen, M. (2017). The SAGE Encyclopedia of Communication Research Methods. The SAGE Encyclopedia of Communication Research Methods

Association of Certified Fraud Examiners(2016). Report to the Nations on Occupational Fraud and Abuse. Retrieved from ACFE: https://www.acfe.com/fraud-resources/report-to-the-nations-archive.

Burmeister, B., & Wermser, H. (2014). Weaknesses of rule-based expert systems and how to overcome them. In Proceedings of the 2014 Conference on Systems Engineering Research (CSER) (pp. 1-5). Redondo Beach, CA, USA: IEEE.

Buku, M., & Mazer, R.(2017). Fraud in Mobile Financial Services: Protecting Consumers, Providers, and the System.

Choudhury, A. et al. (2022). Introduction to digital society: An overview, Advanced Technologies, and Societal Change.

Etherington TR. 2021. Mahalanobis distances for ecological niche modelling and outlier detection: implications of sample size, error, and bias for selecting and parameterizing a multivariate location and scatter method. Peer 9:e11436 https://doi.org/10.7717/peerj.11436

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. Journal of Educational and Behavioral Statistics, 44(3), 348–361. https://doi.org/10.3102/1076998619832248

Himanshu Gore, Rakesh Kumar Singh, Ashutosh Singh, Arnav Pratap Singh, Mohammad Shabaz, Bhupesh Kumar Singh, Vishal Jagota. (2021). Django: Web Development Simple

IMF. (2023, February 22).The IMF and the fight against illicit financial flows. Retrieved from IMF: https://www.imf.org/en/About/Factsheets/Sheets/2023/Fight-against-illicit-financial-flows#

Kang H.(2019).Fraud Detection in Mobile Money Transactions Using Machine Learning.

Kaur, A., Chhabra, J., & Verma, A. (2019). Fraud Detection in Mobile Money Transactions Using Machine Learning Techniques. In Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 68-74).

Liberti, L., & Lavor, C. (2017). Euclidean distance geometry (Vol. 3). Berlin: Springer.

Sánchez-Aguayo, M., Urquiza-Aguiar, L., Estrada-Jiménez, J. (2021).Fraud detection using the fraud triangle theory and data mining techniques: A literature review. Computers, 10(10), 121. https://doi.org/10.3390/computers10100121

Ruankaew, T. (2016). Beyond the fraud diamond. International Journal of Business Management and Economic Research, 7(1), 474-476.

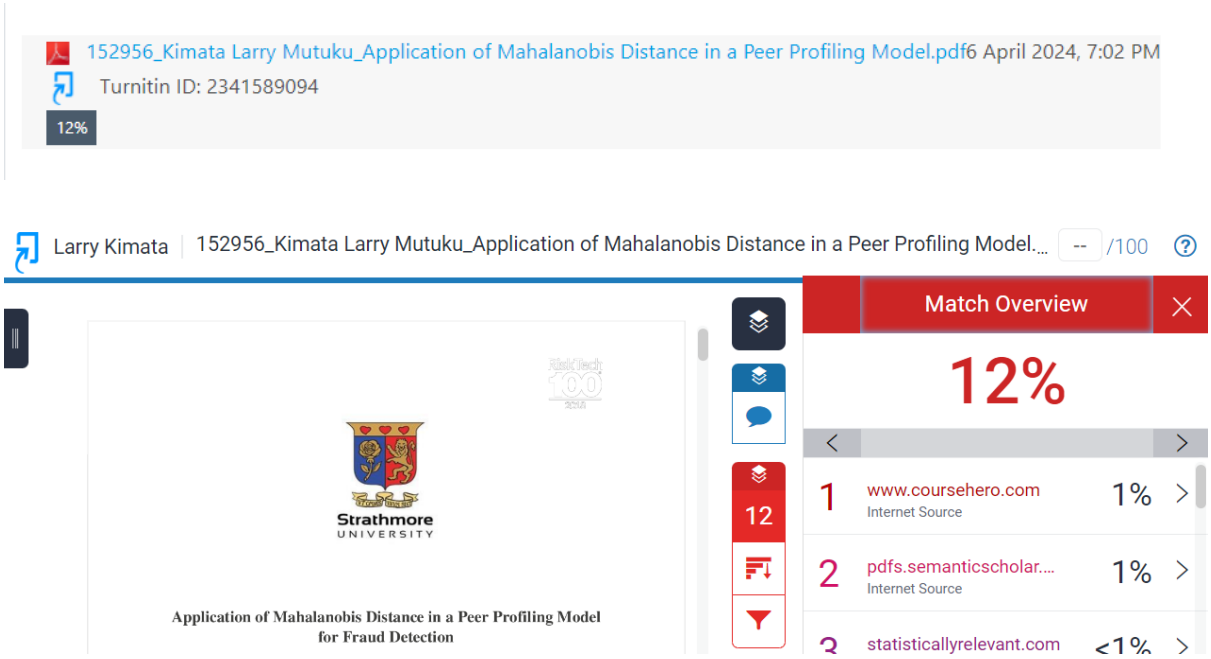
Wells , C. J. (2009, 01 28). *Development Methodologies*. Retrieved 2020, from www.technologyuk.net

Wronka, C. (2022).Cyber-laundering: the change of money laundering in the digital age, Journal of Money Laundering Control, Vol. 25 No. 2, pp. 330-344. https://doi.org/10.1108/JMLC-04-2021-0035

Appendix A: Dataset Sample Portion

step	type	amount	nameOrig	oldbalanceOr	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
1	PAYMENT	9839.64	C1231006815	170136	160296.36	M1979787155	0	0	0
1	PAYMENT	1864.28	C1666544295	21249	19384.72	M2044282225	0	0	0
1	TRANSFER	181	C1305486145	181	0	C553264065	0	0	1
1	CASH_OUT	181	C840083671	181	0	C38997010	21182	0	1
1	PAYMENT	11668.14	C2048537720	41554	29885.86	M1230701703	0	0	0
1	PAYMENT	7817.71	C90045638	53860	46042.29	M573487274	0	0	0
1	PAYMENT	7107.77	C154988899	183195	176087.23	M408069119	0	0	0
1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0	0	0
1	PAYMENT	4024.36	C1265012928	2671	0	M1176932104	0	0	0
1	DEBIT	5337.77	C712410124	41720	36382.23	C195600860	41898	40348.79	0
1	DEBIT	9644.94	C1900366749	4465	0	C997608398	10845	157982.12	0
1	PAYMENT	3099.97	C249177573	20771	17671.03	M2096539129	0	0	0
1	PAYMENT	2560.74	C1648232591	5070	2509.26	M972865270	0	0	0
1	PAYMENT	11633.76	C1716932897	10127	0	M801569151	0	0	0
1	PAYMENT	4098.78	C1026483832	503264	499165.22	M1635378213	0	0	0
1	CASH_OUT	229133.94	C905080434	15325	0	C476402209	5083	51513.44	0
1	PAYMENT	1563.82	C761750706	450	0	M1731217984	0	0	0
1	PAYMENT	1157.86	C1237762639	21156	19998.14	M1877062907	0	0	0
1	PAYMENT	671.64	C2033524545	15123	14451.36	M473053293	0	0	0

Appendix B: Turnitin Originality Report



Appendix C: Consent Form for Participation in the Research Study

Consent Form for Participation on this Research Study

Project Title: Application of Mahalanobis Distance in a Peer Profiling Model for Fraud Detection

Principal Investigator: Larry Mutuku Kimata

Introduction:

You are being invited to participate in a research study for the application of mahalanobis distance in a peer profiling model study.

Purpose of the Study:

The purpose of this study is to explore the challenges faced by financial institutions due to fraudulent activities. We aim to understand and comprehend how these issues affect the fraud landscape to come up with an effective model that solves the fraud issue faced.

Your Involvement:

If you agree to participate, you will be asked to answer a series of questions to assist us understand on the above purpose of study. Your participation is voluntary and free in terms of cost, and you may choose to withdraw or opt out at any item without any issues.

Provisions:

In case one has the difficulty in reading or signing written consent form, we will undertake the consent process verbally using simple and understandable language. One may ask any questions and seek clarification to be able to comprehend well.

Confidentiality:

Any data and information collected during the study will be kept confidential.

Your individual information and data will not be shared with anyone outside of the research team without your explicit permission, except as required by law to provide any information and data for other law purposes.

Benefits and Risks:

While there may not be direct benefits to you personally, this participation to this research study will contribute to the successful outcome of this project which will be key to the fraud landscape and advancements, which may benefit society.

On this research study, there are no risks involved with undertaking this research.

Contact Information:

If you have any questions or concerns about the study, you may contact the Principal Investigator, Larry Mutuku Kimata using the mobile number[+254797071021] or email at [larry.kimata@strathmore.edu] .

If you have any questions or concerns, please do not hesitate to contact me.

Agreement to Participate:

By agreeing to participate in this study, you are indicating that you understand the information provided to you and agree voluntarily to consent to take part on this research study.

Participant's Signature, Email and Date respectively:

---------- ---------- ----------

Appendix D: Interview Form used to achieve Specific Objective one

Interview: Challenges faced by financial institutions due to fraudulent activities

The below entails the interview that will be conducted to achieve the first objective of the research study

Interviewee No: **Position in Fintech /Banking:**

Perception of Fraudulent Activities

Question 1: On a scale of 1 to 10, how do you think and believe that the threat of fraudulent activities is to the financial sector in Kenya?

(1 being not significant at all, 10 being extremely significant)

Question 2: Kindly rate your Kenyan financial institution's vulnerability to fraudulent activities, with 1 being not vulnerable at all and 10 being extremely vulnerable.

Impact Assessment

Question 3: On a scale of 1 to 10, how would you rate the financial impact of fraudulent activities on your Kenyan financial institutions in the last year?

(1 indicating no impact, 10 indicating severe impact)

Challenges in Detection and Prevention

Question 4: On a scale of 1 to 10, how challenging do you find it to keep up with new and evolving types of fraudulent activities?

Question 5: What are some of the fraudulent challenges faced by the financial institutions in Kenya?

Please choose a choice or choices.

A. Lack of a sophisticated fraud tool able to detect.
B. Lack of budget that will assist to acquire a new fraud tool.
C. Lack of knowledge to review the fraudulent transactions.
D. Others, Please Mention it.[……]

Investment in Fraud Prevention

Question 6: Rate your institution's investment in fraud prevention technologies and training on a scale of 1 to 10, with 1 being very underinvested and 10 being highly invested.

Question 7: What are the frauds that impacts at your financial institutions.

A. Account Takeover
B. Love scams
C. Social engineering
D. Bank inside job frauds
E. Others. Please State[……]

Future Outlook

Question 8: On a scale of 1 to 10, how optimistic are you about the future improvements in fraud prevention and detection within the Kenya's financial sector?

Appendix E: SU-ISERC Ethical Approval

Strathmore UNIVERSITY

19th February 2024

Mr Kimata Larry,
larry.kimata@strathmore.edu

Dear Mr Kimata,

RE: Application of Mahalanobis Distance in a Peer Profiling Model for Fraud Detection

This is to inform you that SU-ISERC has reviewed and **approved** your above SU-masters research proposal. Your application reference number is **SU-ISERC1965/24.** The approval period is from **19th February 2024 to 18th February 2025.**
This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
19-Feb-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix F: NACOSTI Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **933508**

Date of Issue: **26/February/2024**

RESEARCH LICENSE

This is to Certify that Mr.. Larry Mutuku Kimata of Strathmore University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: Application of Mahalanobis Distance in a Peer Profiling Model for Fraud Detection for the period ending : 26/February/2025.

License No: **NACOSTI/P/24/33376**

933508

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions