



UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

**USE OF ARTIFICIAL INTELLIGENCE ALGORITHMS TO ENHANCE FRAUD
DETECTION IN THE BANKING INDUSTRY**

By

Edna Mugoshi Shihembetsa

P52/11632/2018

Supervised by Dr. Evans Miriti

AUGUST 2021

Research project report submitted in partial fulfillment for the requirements of the award of the degree of Master of Science in Computational Intelligence, School of Computing and Informatics, University of Nairobi.

DECLARATION

Student

This project report is my original work and has not been presented in any other institution for an academic award. All sources, references, literature used or excerpted during the elaboration of this work are properly cited and listed about the respective source.


SIGNATURE  DATE 31/08/2021

Edna Mugoshi Shihembetsa

Registration Number: P52/11632/2018

Supervisor

This project report has been submitted in partial fulfillment for the requirements of the award of the Degree of Master of Science in Computational Intelligence in the University of Nairobi with my approval as the University Supervisor.

SIGNATURE  DATE 31-08-2021

Dr. Evans A. K. Miriti

ABSTRACT

Fraud is among the most menacing problems with which every human society grapples, given the devastating impact on the effects. This practice refers to the deliberate use of false information to swindle another individual or organization money or property (Association of Certified Fraud Examiners, 2021). The banking industry has for decades used rule-based systems to flag fraud and human review of transactions. Rule-based systems encompass utilizing algorithms which perform a variety of detection actions that are written manually by fraud experts (Oniyilo, 2016). These systems require the manual adjustment of scenarios, which make it challenging to implicitly detect the transactional correlations that would point to fraud. Due to the inherent weaknesses of the rule-based fraud detection approach at banks and limited data that affects commonly used supervised machine learning algorithms, there is an urgent need for new detection techniques or systems that can handle the rapidly increasing fraud and money laundering incidences that adversely affect the Kenyan banking system.

This research aimed to analyse and evaluate various machine learning algorithms to determine their performance in fraud detection for mobile banking transactions within the banking system. The study's objectives were to identify the data attributes that are best suited for mobile banking fraud detection machine learning algorithms and compare the performance of machine learning algorithms in fraud detection in mobile banking transactions.

This study used the CRISP_DM methodology to determine the most accurate fraud detection algorithm. It was published to standardise the data mining processes over the industries. It has since evolved to be the most used methodology in mining data, performing analytics, and projects in data science. Crisp-DM follows the following general steps Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment.

The research results demonstrated that logistic regression did not perform and indicated by the scores and did not predict any fraudulent transactions for the original unbalanced data. As such, it is therefore not recommended for fraud detection. Naïve Bayes performance based on the confusion matrix performed poorly as the algorithm predicted 89,997 false positives, 0 False negatives and 55 True negatives. While predicting the fraudulent transactions accurately, many non-fraudulent transactions were predicted to be fraudulent. These results go hand in hand with results from the scores, which demonstrated that Naïve Bayes had 0.08% accuracy. K Nearest Neighbour had the best results; the algorithm's accuracy was 99%, with 19 true negative predictions, two false positives and 8 False negatives. Therefore, KNN was identified as the preferred algorithm for fraud detection. Additionally, it is noted that when the transactional data is removed KNN, performed marginally better than when the static and scoring data from the fraud detection system are removed.

ACKNOWLEDGEMENTS

I am grateful to God for enabling me to conduct my studies, research, and compile this report. His grace has always been overwhelming.

I am extremely grateful to my supervisor Dr. Evans A. K. Miriti, and the panelists for their continuous guidance, frequent feedback, and valuable time and advice throughout my project.

I am grateful to my family for their support and encouragement during this period of my studies especially my father, Professor Laban Shihembetsa, who instilled the appreciation for education and encouraged the curiosity to learn .

TABLE OF CONTENTS

1. Introduction.....	9
1.1 Background.....	9
1.2 Problem Statement.....	10
1.3 Main Objective.....	10
1.4 Specific Objective.....	11
1.5 Research Questions.....	11
1.6 Significance.....	11
1.7 Justification.....	11
1.8 Scope of Study.....	12
2. LITERATURE REVIEW.....	13
2.1 Fraud Detection.....	13
2.2 Fraud Detection in the Banking Industry.....	13
2.3 Related Work.....	15
2.4 Algorithms Used in Fraud Detection Systems.....	16
2.5 Gaps Identified in Literature Review.....	17
2.6 Description of Proposed Solution.....	18
3. Research Design and Methodology.....	19
3.1 Introduction.....	19
3.2 Research design.....	19
3.3 Business Understanding.....	20
3.4 Data Understanding.....	20
3.5 Data preparation.....	22
3.6 Modelling.....	24
3.7 Experimentation Environment.....	25
3.8 Evaluation.....	26
4. RESULTS AND DISCUSSION.....	28
4.1 Introduction.....	28
4.2 Exploratory Data Analysis.....	28
4.3 Evaluation.....	28
4.4 Working with Unbalanced Data with Logistic Regression.....	31

TABLE OF FIGURES

Figure 1:Proposed Machine Learning Model 18

Figure 2:CRISP_DM Methodology 20

Figure 3:Data Cleansing in Alteryx 23

Figure 4:Masked Data to protect Personal Identifiable Information..... 23

Figure 5:Correlation matrix of Fraud transaction data..... 24

Figure 6:Data Attribute Performance Comparison 33

LIST OF TABLES

Table 1:Confusion Matrix For Logistic Regression, Naive Bayes and KNN	29
Table 2: Score Results for Naive Bayes, Logistic Regression and KNN.....	30
Table 3:KNN Confusion Matrix when N is 5	30
Table 4:KNN Confusion Matrix when N is 2	30
Table 5: Oversampled data Confusion Matrix for Logistic Regression	31
Table 6:Undersampled data Confusion Matrix for Logistic Regression.....	31
Table 7:Confusion Matrix when scored data is dropped.....	32
Table 8:Confusion Matrix when static data is dropped.....	32
Table 9:Confusion Matrix when Transactional data is dropped.....	33

LIST OF ABBREVIATIONS

ML	Machine Learning
KNN	K Nearest Neighbour
BACC	Balanced Accuracy
CRISP_DM	CRoss Industry Standard Process for Data Mining
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative

1. INTRODUCTION

1.1 Background

Fraud is among the most menacing problems with which every human society grapples, given the devastating impact on the effects. This practice refers to the deliberate use of false information to swindle another individual or organization money or property (Association of Certified Fraud Examiners, 2021). The COVID-19 pandemic has seen an increase in fraudulent activities due to the economic downturn as the number of unemployed people during the period increased (Colvin, 2020). Thousands of people were rendered jobless, salaries were reduced, and unemployment rates soared. Naturally, more people have fewer resources at their disposal for their survival, which explains the rise in fraudulent activities as people attempt all means to survive the austere economic times. Fraud experts argue that fraud stems from three elements that include pressure, an opportunity, and rationalisation (Littman, 2011). According to this author, fraud happens when an individual develops an unshareable pressure and motive to commit the fraud.

In most cases, the fraud perpetrator has an unmet need but with limited resources. Unmet needs are endless and vary for different people. It could be a mounting medical bill, reduced income in the household, or gambling debts. Once the person has unmet needs and has limited resources, they identify the opportunity to commit fraud. Perceive opportunities may be reckless management or a lack of internal controls within an entity that would make fraud an easy activity. Lastly, the individual rationalises their decision to commit fraud by convincing themselves that they needed the money more or paying it back eventually. With the tepid economic times, there has been increased pressure and motive to commit fraud, which would make it easy for fraudsters to rationalise their actions.

Fraud is commonplace in the banking industry and includes email phishing, credit card fraud, money laundering, loan application fraud, financial statements fraud, and cyber fraud. With the advent of digital banking, digital fraud has also become more common within this sector. Therefore, it is important to acknowledge that fraud management has become necessary in the banking and commerce industry, which, admittedly, is an excruciating process. Fraudsters have become skillful at discovering loopholes and have established effective techniques such as phishing for unsuspecting individuals and creatively swindling money off them (How Machine Learning Facilitates Fraud Detection? 2021). Therefore, fraud detection methods have to continuously evolve as fraudsters become more effective in designing techniques that bypass rigid banking security systems and learn how to convince unsuspecting individuals to release their money to them.

Traditional Fraud detection methods within the banking industry have been rule-based, where human beings define the rules. 90% of the financial and banking institutions rely on these methods (Onifade and Afolabi, 2015). While more persons adopt new technologies, more fraud scenarios may happen, making those rule-based methods unscalable and unsustainable in the future. Moreover, false positives

(i.e., non-fraudulent transactions catalogued as fraudulent) cause losses in millions of dollars in transactions and customer complaints in the banking industry. Rule-based methods contribute greatly to these outcomes. Ciobanu (2020) conducted a study with 1,000 adult consumers where he found that about 25% of them whose transactions were declined falsely—opted to engage in business with competitors. That rate of switching to competitors increased to 36% for consumers aged between 18 and 24 years old. It also increased to 31% for those aged between 25 and 34 years old. These study results indicate the dire need for more rigorous and modern fraud detection methods.

To add to the challenge with the traditional rule-based system, fraudsters lack specific patterns and constantly change their behaviour over time, making the systems cumbersome and rapidly obsolete. There is a clear need for a change of approach in security systems within the banking systems. According to the Nilson report (2019), it was anticipated that fraud involving cards only amounted to a staggering amount of \$30 billion globally, by 2020. Additionally, with the technology disruption within the banking sector because of the existence of numerous payment channels such as credit and debit cards, smartphones, the rate of transactions has exponentially increased over the past few years. Fraudsters have also developed extremely effective fraudulent tactics. Given this situation, there is a need to develop more rigid and robust fraud detection approaches in banks. The most viable option is machine learning algorithms installed in banking systems.

1.2 Problem Statement

The banking industry has for decades used rule-based systems to flag fraud and human review of transactions. Rule-based systems encompass utilizing algorithms which perform a variety of detection actions that are written manually by fraud experts (Oniyilo, 2016). These systems require the manual adjustment of scenarios, which make it challenging to implicitly detect the transactional correlations that would point to fraud. Due to the inherent weaknesses of the rule-based fraud detection approach at banks and limited data that affects commonly used supervised machine learning algorithms, there is an urgent need for new detection techniques or systems that can handle the rapidly increasing fraud and money laundering incidences that adversely affect the Kenyan banking system.

1.3 Main Objective

This research aims to analyse and evaluate various machine learning algorithms to determine their performance in fraud detection for mobile banking transactions within the banking system. There is a need for real-time fraud detection methods to help banks protect themselves and their customers as transactions happen in real-time.

1.4 Specific Objective

1. To identify the data attributes that are best suited for mobile banking fraud detection machine learning algorithms.
2. To compare the performance of machine learning algorithms in fraud detection in mobile banking transactions.

1.5 Research Questions

1. What are fraud detection scenarios currently being used in traditional fraud detections systems?
2. Which machine algorithms are suitable for fraud detection for mobile banking transactions?
3. What features in machine algorithms are suitable for fraud detection?
4. What is the performance of the machine learning algorithms selected for mobile banking transaction fraud detection?

1.6 Significance

The significance of the study is the addition to the growing body of research on machine learning for mobile banking fraud detection in the Kenyan banking industry. The resulting product will aid in the development of automatic fraud detection systems without human intervention. This study demonstrates how to identify an algorithm that will reduce false positives and preserve the banks' customers.

1.7 Justification

An effective fraud detection system should accurately detect the transaction in real-time. There are two types of fraud detection systems that include anomaly detection and misuse detection. Anomaly detection systems detect intrusions into systems and uncover any outliers in the data by monitoring the system activities. Misuse detection systems detect attacks within the systems (Aghaei, 2017). This is achieved by defining normal behaviour within the system and then setting all other behaviour as abnormal, where it is flagged in real-time within the misuse detection systems.

Several approaches in machine learning (ML) have been implemented over the years. Typical ML algorithms used are KNN (K Nearest Neighbour), decision trees, and Logistic Regression. However, these are supervised methods, implying that they need to learn by labels to identify fraudulent transactions. When the company lacks this information, these algorithms are untrainable.

Given the increase in fraudulent activities within the banking sector, this research project must explore machine learning algorithms' effectiveness, including KNN (K Nearest Neighbour), Naïve Bayes and logistic regression over traditional rule-based fraud detection systems. The loopholes within the rule-based systems necessitate more automated fraud detection approaches.

1.8 Scope of Study

The analysis focused on Equity Bank Limited transactional data. The data consisted of data from the mobile banking system that included data from mobile banking transactions and data from the fraud detection system. The dataset extracted from the data warehouse for this research comprised 450,352 separate mobile banking transactions for February 2021. It is important to note that the prototype solution presented in the paper can be applied to other banks. This prototype is applicable where the type of data captured is similar to that used in creating the model.

2. LITERATURE REVIEW

2.1 Fraud Detection

More often than not, there is confusion between bank fraud and bank robbery. Bank robbery differs from bank fraud, where bank robbery involves violence, while bank fraud is often a slow and secretive process that goes undetected until the deed is complete. Bank fraud often entails coaxing the customers to provide their sensitive bank details, which may be used to obtain money remotely.

Fraud detection entails activities undertaken avoid obtaining money or property through false fabrications. Within the banking sector, fraud entails check forgery and the utilization of stolen credit card. It may entail the exaggeration of losses or creating unfortunate events such as accidents with the sole intent for the pay-out (Rouse, 2019).

According to the Nilson Report (2019), it was anticipated that fraud involving cards only amounted to a staggering amount of \$30 billion globally, by 2020. Additionally, the disruption in technology in banking and payments because of an increase in payment channels such as credit and debit cards and smartphones, the number of transactions has significantly increased over the recent years.

Fraud detection efforts using data analytics, software designed for fraud detection, and tools, programs designed for fraud detection enable organisations to foresee common fraud tactics, automate the process of cross-referencing of data, continuously monitor their transactions in real-time, and discover the latest and sophisticated fraudulent schemes in the sector (Association of Certified Fraud Examiners, 2021). Fraud detection and prevention resources such as software are available in both proprietary and free versions. Some of the common features in this software include a dashboard, data importation and exportation, visualisation of data, the integration of customer relationship management, managing calendar, scheduling, budgeting, and password management. They also have Application Programming Interfaces (API), billing, two-factor authentication, and customer database management.

2.2 Fraud Detection in the Banking Industry

According to an article on Javelin Strategy & Research on fraud detection, it takes longer for fraud detection in brick-and-mortar financial organizations (Pascual et al., 2017). This amount of time is unfavourable for the customers and financial institutions because multiple fraudulent activities could occur before detection within such a time frame. Fraud adversely affects banks that deal with online payments services, especially within the contemporary technological advancement in the business sector. For example, about 20% of clients shift banks after experiencing fraud (Sando, 2021). This number of customers leaving a particular bank is significantly detrimental to a bank's business operations, especially if the trend continues over several years. There is a dire need for financial institutions to establish proper and robust fraud detection approaches within their systems to curb this

act. Two major fraud detection approaches exist, which include the rule-based approach and machine learning fraud detection.

2.2.1 The Rule-Based Approach.

Activities involving fraud within the finance realm are detectable by exploring on-surface and clear signals. Despite it being unusual, large transactions and those that occur in uncommon place should undergo more robust verification. Rule-based systems utilize algorithms that assess various fraud detection situations which are written manually by fraud analysts. Currently, legal systems use approximately 300 various rules approve transactions. This explains why using the rule-based systems is a straightforward process. These systems require manually adjusting scenarios and situations which are unable to detect the implicit correlations. Additionally, rule-based systems often use outdated software which hardly process real-time data streaming through the systems that is important in the digital sphere (Moon et al., 2017).

The rules-based systems depend on the pre-determined rules so as to recognize variations in behaviour. These systems are rigid and unable to adapt to industries such as financial services that require a more lithe and agile platforms to overcome challenges associated with identifying fraud (Moon et al., 2017). Rule-based approaches are time-intensive as they require fraud analysts to establish the rules used to detect fraud. These approaches also require manual work and multiple verification processes that obstruct user experience. To add to this, rules-based approaches identify obvious fraud patterns and are therefore not adaptable to the changes in fraud that occur as fraudsters evolve. On the other hand, rule-based models become more costly in maintaining the data set or the customer base size expands.

2.2.2 Machine Learning-Based Fraud Detection

There exist concealed events in user behaviour that may not be outrightly evident showcase possible fraudulent transactions. Machine learning permits the creation of algorithms that can process bigger datasets with several variables and helps detect these concealed correlations between operator behaviour and the possibility of fraudulent activities. Machine learning systems are better than rule-based systems because they are quicker in data processing and are less manual in handling. For instance, smart algorithms are congruent with behaviour analytics in the reduction of verification steps required.

Firms that deal with regulation of financial services are involved in the monitoring likely fraudulent activity: where they must detect communicate to each other about the flagged activities. Villalobos et al. (2019) describe an instance in which a machine learning prototype was programmed on a dataset that had transactions criminally completed. The prototype used with the rule-based system helped discover concealed relations between the transactions and criminal activities. Such systems minimise the workload in the smaller banks involved in fraud monitoring. The proposed solution showed that 99.6% of money laundering transactions and reduced the reported transactions from 30% to 1%.

ML depends upon algorithms, that are more effective as the size of the data sets increase. The more the data, the more ML prototype improves and can distinguish the similarities and differences over various behaviours. The more the ML model identifies the difference between the legitimate and fraudulent transactions, the more the systems become efficient in sorting out the data sets into various categories. ML systems are therefore more scalable as the customer database grows.

While ML algorithms present numerous benefits, they have significant drawbacks that limit their use in fraud detection. For instance, one of the drawbacks is that ML requires significant amount of data for the models to achieve accuracy. This data volume is manageable; however, there should be sufficient data points that recognize the legitimate causal relationships in smaller organisations. Additionally, machine learning models function on actions, behaviour, and activity. The model tends to overlook clear connections, such a card used on two different accounts, hence, rendering the fraud detection activity ineffective.

2.3 Related Work

Ayowemi et al. (2017) researched fraud detection related to credit cards as this becomes an impediment because of two major reasons. Firstly, behaviour profiles that are normal and fraudulent constantly change, and second, the data sets from credit card fraud are often unbalance. The approach used in sampling on the dataset, how the variables are selected, and the techniques used in detection greatly affect the fraud detection performance in credit card transactions. The researchers investigated the performance of k-nearest Neighbour, naïve Bayes, and logistic Regression on credit card data considered highly skewed. A hybrid technique where the skewed data was under-sampled and oversampled was carried out. These techniques were used on the data and later applied in Python. The technique performance was evaluated based on factors such as their accuracy, specificity, sensitivity, Matthew's correlation coefficient, precision, and the balanced cataloguing rate. The results showed higher accuracy levels for naïve Bayes, k-nearest Neighbour, and classifiers on Logistic Regression were 97.92% and 97.69%. The comparative results indicated that the k-nearest neighbour performed better than the Logistic Regression and naïve Bayes algorithms.

Bauder et al. (2017) studied claimed fraud in Medicare. The researchers compared various ML methods used in the detection of Medicare fraud. They performed a comparative study with hybrid machine learning methods based on 4 performance systems of measurement and reduction of class imbalance by oversampling and using the 80-20 under-sampling approach. Their results indicated that there was a bigger gap in performance gap between the methods used in sampling. The latter sampling approach had better novice performance compared to former sampling approach, which is the oversampling. Their research noted that oversampling indicates poor performance for all machine learners.

To add to this, Balanced accuracy (BACC) was considered undependable in measuring performance of models across all methods and incapable of sufficiently reflecting the more realistic alterations

perceived in the other metrics. It is safe to claim that the learner performance was improved following the under-sampling approach with the supervised methods being pointedly better than the unsupervised and hybrid machine learners. Finally, the provider category contributed to the challenges in fraud detection with fairly specialised provider categories demonstrating better performance over other general specialities.

While research has been done on fraud detection using machine learning, these studies have focused on specific types of transactions, especially credit card fraud. No studies have focused on all types of transactions in a bank, as fraud does not occur in isolation. To effectively detect fraud, there is a need to look at the full range of processed transactions in a bank. This will enable fraud detection systems to detect anomalies across different transactions, therefore increasing detection accuracy.

2.4 Algorithms Used in Fraud Detection Systems

Algorithms in machine learning are designed to learn by themselves using experience.

2.4.1 Using Logistic Regression for Fraud Detection Machine Learning Algorithms

Logistic regression refers to a supervised learning method used with definite decisions. This means that the results obtained are considered fraud or non-fraud in the event of a transaction. This approach utilizes a cause and effect relationship to develop organized data sets. Regression analysis technique is more sophisticated when utilized in detection of fraud because of the several variables and data set sizes. This model (algorithm) predicts whether new transactions are flagged as fraudulent. These models are usually precise to their clients from larger merchants, but usually, general models remain applicable.

2.4.2 Using Decision Tree for Fraud Detection Machine Learning Algorithms

Decision Tree algorithms are used to classify atypical activities in a transaction from an authorised user. These algorithms have trained constraints used on the dataset in fraud classification. Decision Tree algorithms are utilized in the classification or regression extrapolative modelling difficulties. They are fundamentally a rule sets skilled to use fraud cases involving clients.

Creating a decision tree disregards unrelated features and does not necessitate wide-ranging data normalisation. Once a tree undergoes inspection, it is understood the reason why a specific decision was made by depending on the list of rules activated by a specific client. The machine learning algorithm output may be a model that apes the decision tree. This gives a possibility score of fraud based on earlier circumstances set.

2.4.3 Using Random Forest for Fraud Detection Machine Learning Algorithms

Random Forest ML uses a mixture of decision trees to enhance the results. Each tree assesses the transactions for different conditions (Ayyadevara, 2018). Random datasets are trained. Depending on the decision trees training, each tree provides classification of a transactions as fraudulent or non-fraudulent. Then, the model is used to accordingly predict the result. It allows fraud detectors to even out the error that may be present in a tree. It enhances the general performance model accuracy while sustaining the ability to interpret the results and provide explicable scores to our users.

Random forest runtimes are quick and handle data that is missing or unbalanced. Random Forest MLs have weaknesses such as; when used in regression, they are unable to predict beyond the variety in the training of the data, and they may over-fit data sets considered noisy.

2.4.4 Using Neural Networks for Fraud Detection Machine Learning Algorithms

Neural Networks is based on the human brain. They use different computational layers. They utilize cognitive computing that helps build machines that can use self-learning algorithms that include data mining, recognizing patterns, and process natural language (Graupe, 2016). Neural networks undergo several layers for the data training process. It provides more precise outcomes compared to other models as it uses cognitive computing and learns from the patterns of authorised behaviour; thus, it distinguishes between 'fraud' and non-fraud transactions. The neural networks are wholly adaptive and learn from patterns set of legitimate behaviour. These acclimatize to the alteration in the behaviour of what are considered standard transactions and identify forms of fraud transactions. The neural networks process is fast and operates in real-time.

2.5 Gaps Identified in Literature Review

While extensive work has been done in fraud detection for financial services industry, banking, and insurance, the research inadequately covers mobile banking transactions focusing on the Kenyan environment. Additionally, the research done has been done based on the European market, which has different transaction patterns from the Kenyan demographic. Therefore, research is needed to focus on mobile banking transactions and the data captured by the bank's systems to determine how the algorithms perform on such transactions.

2.6 Description of Proposed Solution

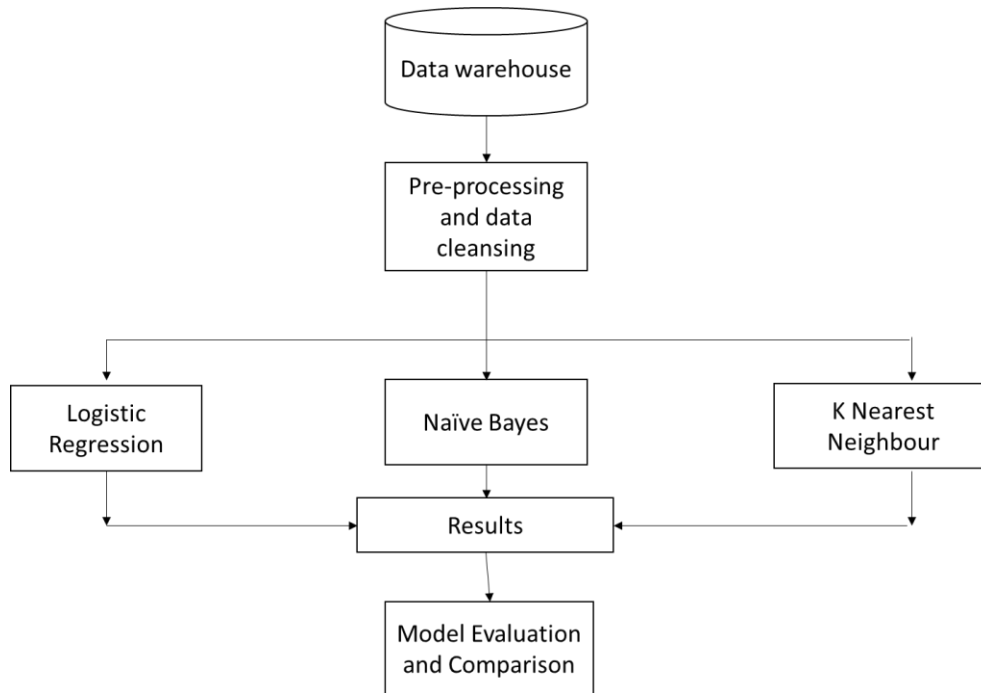


Figure 1: Proposed Machine Learning Model

Fraud detection is a method to identify fraudulent transactions as they are performed in the banking system. The goal is to identify the fraud before it leads to monetary and reputational loss to the bank to keep customers satisfied with the safety of their deposits. The proposed solution will involve mobile banking transactions with customer details (age, gender, location) to develop the classifiers using three algorithms, Naïve Bayes, Logistic Regression, and K nearest Neighbour. Results from the three algorithms will be analysed and compared to identify the most accurate algorithm.

The choice of these algorithms was based on related work that demonstrated promising results with Naïve Bayes and K nearest Neighbour when tested for accuracy. Logistic regression has shown some promising results, and with a larger data set, the algorithm's accuracy will increase.

CHAPTER 3 : RESEARCH DESIGN AND METHODOLOGY

3.1 Introduction

The chapter highlights how the study on machine learning for fraud detection has been structured. It covers the study's research design, the data collection process, model development, and evaluation metrics.

3.2 Research design

This is described as a framework of techniques and approaches identified to combine different research attributes by a researcher to handle the research problem logically. This process helps the researcher establish the plan for gathering, analysing, and evaluating the collected data. It is pertinent to have a guideline on how the research questions and the research objectives would be responded to and how the tracking of stages of the research will be done. It evaluates the research purpose, methods, and approaches, and time limit. Thus, the research design answers questions on what data needs to be collected and how data collection and analysis should be done.

Research designs for the quantitative approach are descriptive, where subjects are measured only once or experimental manner. In this study, we will use the CRISP_DM method to determine the most accurate fraud detection algorithm.

CRISP DM_Data Science

CRISP DM stands for Cross Industry Standard Process for Data Mining which is a phased process model that logically describes the data science life cycle (Kapoor, 2019). These are a set of steps that assist in planning, organising, and implementing a data science (or machine learning) project.

The methodology was published as a way to standardise the process of data mining across various industries and is now the most frequently adopted methodology in data analytics, data mining, and data science projects.

Crisp-DM follows the following general steps:

- i. Business understanding
- ii. Data understanding
- iii. Data preparation
- iv. Modelling
- v. Evaluation
- vi. Deployment

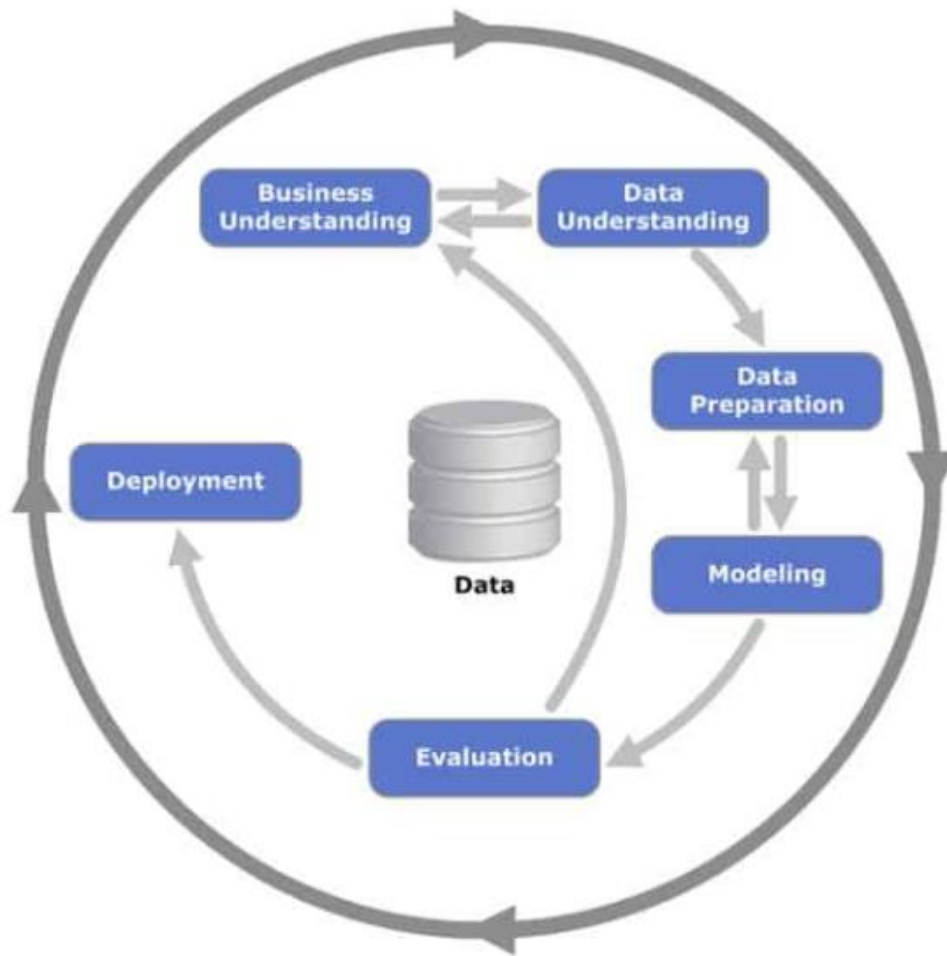


Figure 2:CRISP_DM Methodology

Phases in Crisp-DM Methodology

3.3 Business Understanding

The bank has witnessed an increase in fraudulent mobile banking transactions which has led to losses by both customers and the bank. The business, therefore, requires a fraud detection system that will accurately flag transactions in real time.

3.4 Data Understanding

The research data was extracted from Equity bank, a Tier 1 bank in Kenya serving over 14 million customers who perform millions of transactions in a day. The bank stores data from all major systems in a Data warehouse. The data consisted of data from the mobile banking system that includes data from mobile banking transactions and data from the fraud detection system. The transaction data in the data warehouse was combined with the list of known reported fraud cases to generate markers on the data for fraudulent and non-fraudulent transactions.

The dataset extracted from the data warehouse for this research comprised 450,352 distinct mobile banking transactions for February 2021. The dataset contained the following fields:

- 1) TRAN_DATE - DateTime
- 2) VALUE_DATE - DateTime
- 3) TRAN_ID - String
- 4) ACID - String
- 5) CUST_ID - String
- 6) FORACID - String
- 7) ACCT_NAME - V_String
- 8) SCHM_CODE - String
- 9) TRAN_PARTICULAR - V_String
- 10) TRAN_CRNCY_CODE - String
- 11) DELIVERY_CHANNEL_ID - String
- 12) TRAN_AMT - Double
- 13) PART_TRAN_TYPE - String
- 14) ENTRY_DATE - DateTime
- 15) PSTD_DATE - DateTime
- 16) VFD_DATE - DateTime
- 17) REF_NUM - String
- 18) REF_AMT - Double
- 19) TRAN_PARTICULAR_CODE -String

Analysis of the records showed that the data had the following errors in the fields

- 1) TRAN_ID – 100% of the records had leading or trailing whitespaces
- 2) REF_NUM- 24% of the data had empty values, and 0.01 % of the data has whitespaces
- 3) TRAN_PARTICULAR_CODE- 90.80% of the data has empty values

In addition to the mobile banking data, the research required fraud transactions detected by the rule-based fraud detection system. The dataset extracted from this system contained 274 unique fraudulent transactions with the following attributes:

- 1) TRAN_DATE - DateTime
- 2) TRAN_ID - String
- 3) CUST_ID - String
- 4) TRAN_PARTICULAR -V_String
- 5) TRAN_AMT - Double
- 6) IS_FRAUD - V_WString
- 7) AGE_SCORE - V_WString
- 8) FIRST_TIME_SCORE - V_WString
- 9) SIM_SWAP_SCORE - V_WString

A review of the data showed that the dataset showed that the column TRAN_ID contained leading and trailing whitespaces. It was, however, noted that the other column's data had no errors.

3.5 Data preparation

To prepare the data for the modelling the Alteryx application which is a data analytics software equipped with handling large data sets was used. It can perform complex analyses on data.

1. To create the master data, the two datasets were concatenated using TRAN_ID as the Primary Key to combine the fields. This resulted in a combined dataset with 23 unique columns which were: TRAN_DATE, TRAN_ID, CUST_ID, TRAN_PARTICULAR, TRAN_AMT, IS_FRAUD, AGE_SCORE, FIRST_TIME_SCORE, SIM_SWAP_SCORE, VALUE_DATE, ACID, FORACID, ACCT_NAME, SCHM_CODE, TRAN_CRNCY_CODE, DELIVERY_CHANNEL_ID, PART_TRAN_TYPE, ENTRY_DATE, PSTD_DATE, VFD_DATE, REF_NUM, REF_AMT and TRAN_PARTICULAR_CODE.

2. The next step was to confirm that the correct data types were assigned to the fields, including dates that were stored as strings were converted to date

Auto Field (11) The FieldType of "TRAN_ID" changed to: String(9)

Auto Field (11) The FieldType of "CUST_ID" changed to: String(9)

Auto Field (11) The FieldType of "TRAN_PARTICULAR" changed to: String(50)

Auto Field (11) The FieldType of "IS_FRAUD" changed to: Byte

Auto Field (11) The FieldType of "AGE_SCORE" changed to: Byte

Auto Field (11) The FieldType of "FIRST_TIME_SCORE" changed to: Byte

Auto Field (11) The FieldType of "SIM_SWAP_SCORE" changed to: Byte

Auto Field (11) The FieldType of "ACID" changed to: String(10)

Auto Field (11) The FieldType of "FORACID" changed to: String(13)

Auto Field (11) The FieldType of "ACCT_NAME" changed to: V_String(80)

Auto Field (11) The FieldType of "SCHM_CODE" changed to: String(5)

Auto Field (11) The FieldType of "TRAN_CRNCY_CODE" changed to: String(3)

Auto Field (11) The FieldType of "DELIVERY_CHANNEL_ID" changed to: String(6)

Auto Field (11) The FieldType of "PART_TRAN_TYPE" changed to: String(1)

Auto Field (11) The FieldType of "REF_NUM" changed to: V_String(20)

Auto Field (11) The FieldType of "TRAN_PARTICULAR_CODE" changed to: String(3)

3. Replaced all blank fields with null where the data type was string.
4. Replaced all blank fields with 0 where the data type was numeric.
5. Removed whitespaces from the data, including leading and trailing whitespaces and space, tab and duplicate white spaces.

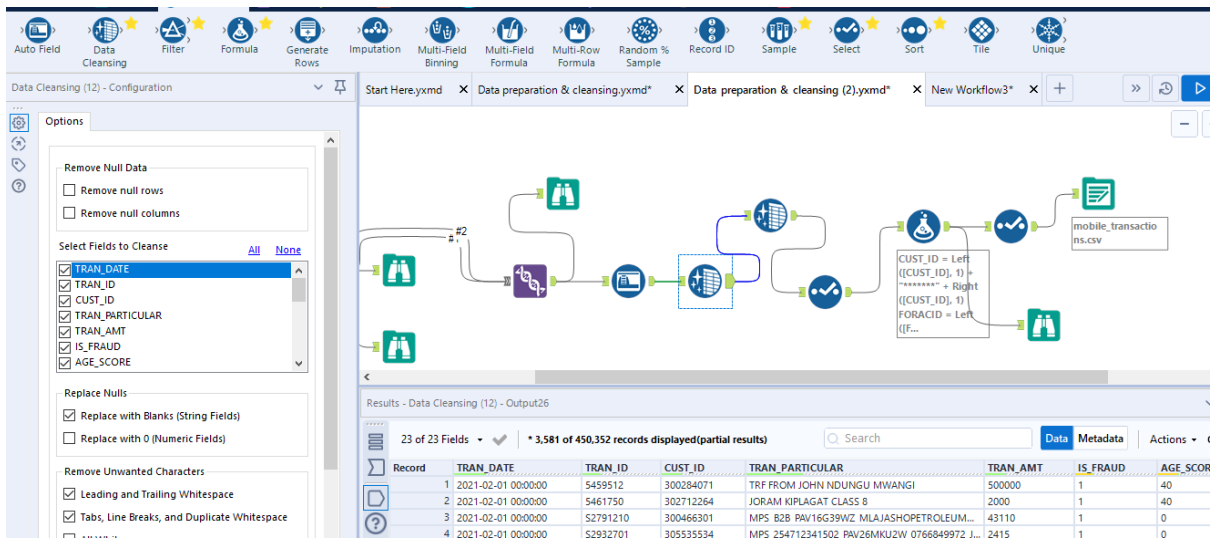


Figure 3: Data Cleansing in Alteryx

- Due to the joining the mobile transactions data with the fraud data had four extra columns that were not available in the original mobile banking data set. This required that all other records have a value included in those fields. The data in the blanks created was replaced by 0 as they are numerical fields.
- As the data contains sensitive customer records the fields containing any Personally identifiable information were masked to maintain the confidentiality of the customers.

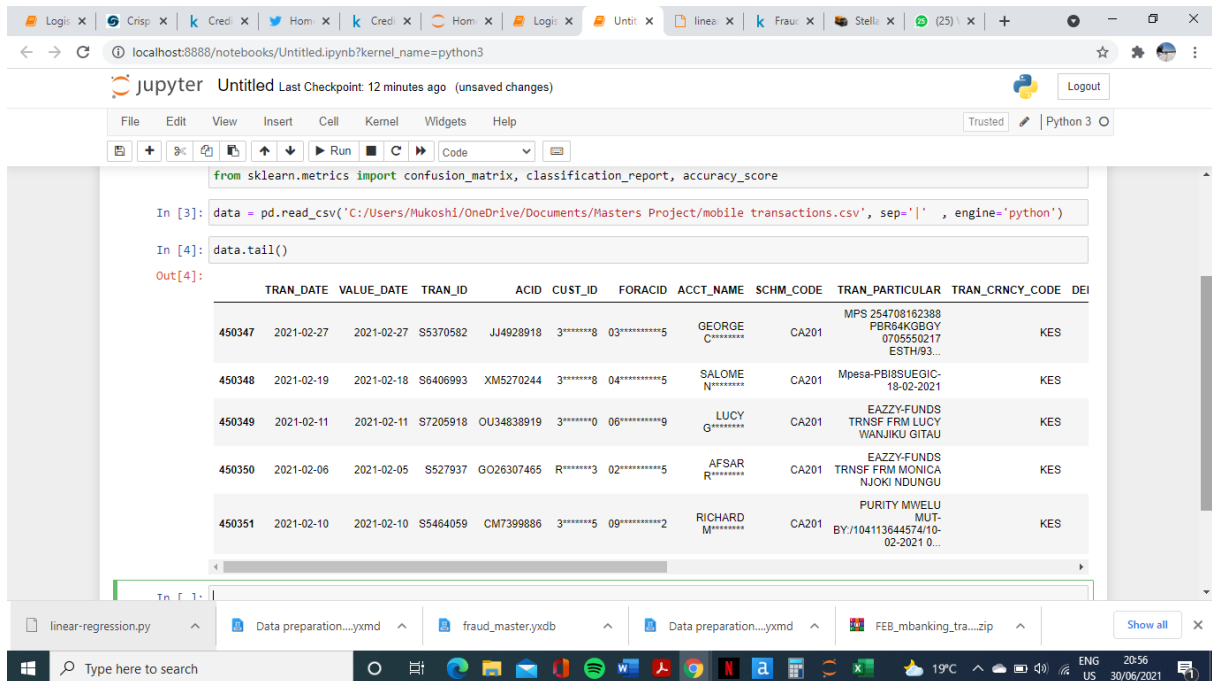


Figure 4: Masked Data to protect Personal Identifiable Information

To further understand the data, below is a correlation matrix :

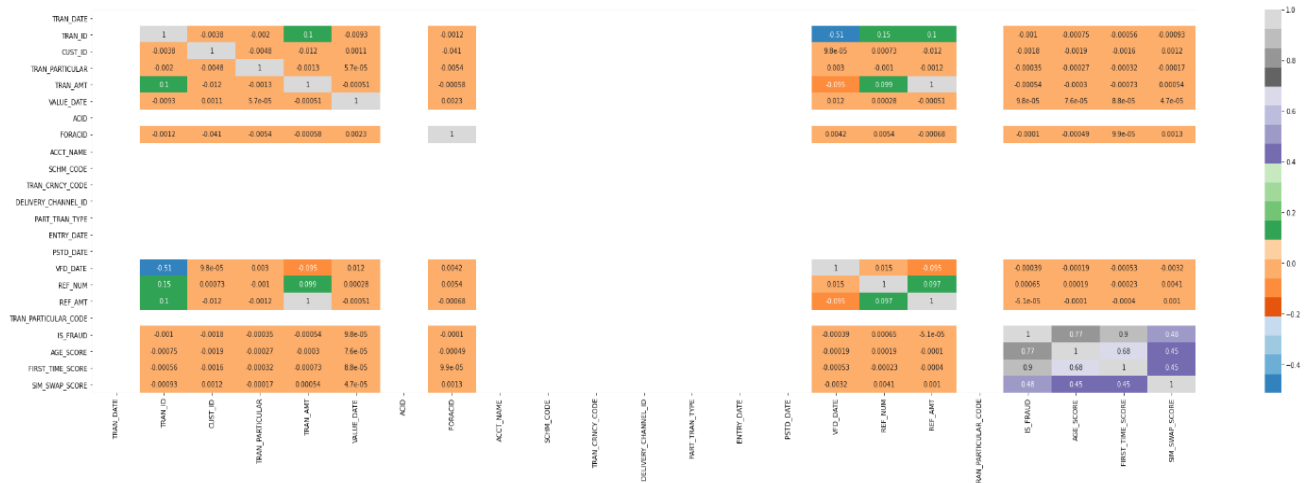


Figure 5:Correlation matrix of Fraud transaction data

3.6 Modelling

As discussed in previous chapters, the algorithms selected for this study were Logistic Regression, Naïve Bayes and KNN. The dataset was split, with 80% of the data being for training the model and 20% for testing the model. A review of the data showed that the data was unbalanced as there were only 274 fraudulent transactions in 450,352 mobile transactions.

Logistic Regression

Logistic regression algorithm utilises a statistical function to approximate the likelihood of a binary feedback based on the variables. It identifies the parameters that are best suited to a non-linear function which is called the sigmoid.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$x = w_0 z_0 + w_1 z_1 + \dots + w_n z_n$$

The vector labelled (z) which is input data, and the most suited coefficients were multiplied with every element and added up which then results in one number. This number determines the target class's classification. If the sigmoid value is greater than 0.5, it is considered as 1, that is fraudulent; otherwise, it's a 0, non fraudulent. Stochastic gradient ascent progressively appraises the classifier as the new data is updated. The weight is set to 1 for all at the beginning.

The gradient ascent was calculated for all feature values that are in the dataset. The weights vector is updated by the product of gradient and alpha and the result are the weight vector. Stochastic gradient ascent was utilised in this research as the dataset was large; the weights are updated by using only one instance at a time, therefore, resulting in reduction of complexity(Awoyemi et al.,2017).

Naïve Bayes Classifier

This is a machine learning algorithm that uses the statistical approach based on Naïve Bayes theory. The theory selects the choice based on highest likelihood. Bayesian probability approximates unknown likelihoods from known values. The algorithm allows prior logic and knowledge to be used to uncertain statements. This approach has a presupposition of conditional independence in data features (Awoye mi et al., 2017).

$$P(c_i | f_k) = \frac{P(f_k | c_i) * P(c_i)}{P(f_k)}$$
$$P(f_k | c_i) = \prod_{i=1}^n P(f_k | c_i) \quad k=1, \dots, n; i=1, 2$$

where n is a representation of the maximum count of features, while, $P(c_i | f_k)$ is the likelihood of the feature value f_k being in the class c_i , while $P(f_k | c_i)$ is likelihood of generating feature value f_k given class c_i , $P(c_i)$ and $P(f_k)$ are likelihood of class c_i occurring and likelihood of feature value f_k occurring sequentially. The Binary classification is performed based on Naïve Bayes Classification.

If $P(c_1 | f_k) > P(c_2 | f_k)$ then the classification is C_1

If $P(c_1 | f_k) < P(c_2 | f_k)$ then the classification is C_2

K Nearest Neighbour

In KNN, an incoming transaction is classified by computing the nearest neighbour to the incoming transaction. In the likelihood the nearest neighbour is a fraudulent transaction, the transaction then is indicated as a fraudulent transaction. The value of K utilised is odd and small to break the any ties (1, 3 or 5 are typically used). A larger value of K can assist to reduce the effect of noisy a data set. In KNN, the distance between two data occurrences can be calculated in different ways. When the attributes are continuous, Euclidean distance would be considered the ideal choice (Harrison, 2021). A simple matching coefficient is preferred in the case of categorical attributes,. The distance is usually computed for each attribute and then combined for multivariate data. For performance improvement of the KNN algorithm the distance metric can be optimised. This technique requires both fraudulent and non- fraudulent transactions.

3.7 Experimentation Environment

The preferred environment to perform the experimentation was Anaconda Distribution, an open-source machine learning environment that provides an efficient platform for creating data science python or R scripts on all operating systems.

3.8 Evaluation

After the training, to assess the model's functionality, we introduced to the model completely new data, for which we were aware of the fraudulent transactions. If the model detected the fraud accurately, it can be deployed against mobile banking transactions.

To determine the right risk threshold data analysis was performed based on the principles of precision and recall. It is balancing act between the following values:

- True positives(TP) (how many fraudulent transactions we block)
- False positives(FP) (how many non-fraudulent transactions we block)
- False negatives (FN)(how many fraudulent transactions we allow)
- True negative(TN) (how many non fraudulent transactions we allow)

Accuracy

Is described as the ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Precision

Precision is described as the ratio of accurately predicted positive observations to the aggregate predicted positive observations.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

F1

The weighted average of Precision and Recall results in the F1 score. This score ,therefore, takes both false negatives and false positives into account.

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

Recall(Sensitivity)

Recall is the ratio of accurately predicted positive observations to the total observations in actual class.

Recall = TP

TP+FN

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter discusses the outcomes of this research concerning the main objective: to analyse and evaluate various machine learning algorithms to determine their performance compared to rule-based systems, as discussed in chapter one. Three machine learning algorithms, KNN (K Nearest Neighbour), Naïve Bayes and Logistic Regression, were tested iteratively, evaluating different sets of variables to determine the best fraud detection outcomes.

4.2 Exploratory Data Analysis

The dataset was sourced from the Equity Bank data warehouse. The dataset contains mobile banking transactions made by Equity bank customers in February 2021. This dataset contains mobile banking transactions were transacted in one month, with a total number of 450,352 mobile banking transactions. The positive class which are the fraud transactions constitute 0.0609% of the transactions data. The dataset is skewed towards the fraudulent transactions and unbalanced. The data contains numerical variables and date variables. The details of the transactions and background meta data of the features could not be extracted as this was confidential data. The feature 'IS_FRAUD' is the target class for the binary categorization, and the values are one(1) for positive case (fraud) and zero(0) for negative case (non-fraud).

4.3 Evaluation

K-NEAREST NEIGHBOUR

K-nearest neighbour is a machine learning algorithm that is instance-based which computes the classification of fraudulent transactions based on a similarity measure, such as Minkowski, Euclidean, or Manhattan distance functions. Euclidean and Manhattan distance measures are best suited with continuous variables, while the Minkowski is suited categorical variables (Awoyemi et al., 2017). For this research the Euclidean distance measure was used for the classifier.

To calculate the Euclidean distance between two input vectors the equation is as given below:

$$D_{ij} = \sqrt{\sum_{k=1}^n (X_{ik} - X_{jk})^2} \quad k=1,2,\dots,n$$

For each data point, a computation of the interval between a random input data point and the current point is performed. The distances are then sorted in ascending order and k items with the lowest distances to the input data point. Once the majority class is found, the categorisation algorithm gives back the majority class as the class for the input point (Awoyemi et al., 2017).

Logistic Regression

The logistic regression model calculates a weighted sum of input variables and bias, and doesn't output the outcome, instead it undergoes a logistic function. If the likelihood is calculated to be more than 50%, then the model predicts that the instance belongs to that class; otherwise, it does not belong to that class. It is essentially considered a binary classifier (Brownlee, 2021).

Naïve Bayes

Naïve Bayes is a statistical approach that is based on the Bayesian theory. This approach selects the decision as a result of highest probability and estimates unknown likelihood from values that are known. One can apply previous information and logic to undetermine statements.

PERFORMANCE EVALUATION AND RESULTS

TP(True Positive) are transactions classified as positive correctly. TN(True negatives) are transactions classified correctly as negative. False-positives are transactions classified as positive but are actually negative transactions. While false-negatives are transactions classified as negative however, they are positive. The performance of the three classifiers were evaluated based on the accuracy, recall, precision and F1 scores (Awoyemi et al.,2017).

For the study, three different classifier models were developed based on KNN ,Logistic Regression and Naïve Bayes. For evaluation of these models, 80% of the dataset was utilised for training while 20% is set used for validation and testing for KNN ,Logistic Regression and Naïve Bayes,.

To further understand the results from the data, we reviewed the confusion matrix and the results demonstrated as below:

	True Negative	False Positive	False Negative	True Positive
Logistic Regression	90018	0	53	0
Naïve Bayes	19	89,997	0	55
K-nearest Neighbour	47005	2	8	19

Table 1:Confusion Matrix For Logistic Regression, Naive Bayes and KNN

The confusion matrix revealed that logistic Regression did not perform as well as indicated by the scores and did not predict any true positive values. That means it did not predict any fraudulent transactions at all. The high accuracy, F1, Recall and Precision scores were a result of division by 0. This, therefore, means that logistic Regression does not perform well for unbalanced data and as fraud data is inherently unbalanced. As such, it is therefore not recommended for fraud detection.

Naïve Bayes performance based on the confusion matrix performed worse than logistic as the algorithm predicted 89,997 false positives, 0 False negatives and 55 True negatives. These results go hand in hand with results from the scores, which demonstrated that Naïve Bayes had 0.08% accuracy. K Nearest

Neighbour had the best results; the algorithm's accuracy was 99%, with 19 true negative predictions, two false positives and 8 False negatives. In comparison, K nearest Neighbour outperformed the other algorithms.

	Accuracy	Precision	Recall	F1
Logistic Regression	0.999	1	1	1
Naïve Bayes	0.00082	0.00061	1	0.0012
K-nearest Neighbour	0.99977	0.99977	0.7037	0.79166

Table 2: Score Results for Naive Bayes, Logistic Regression and KNN

To determine the best value of N for KNN, tested the algorithm when N was set to 2 and when it was set to 5. Below are the results when the Nearest Neighbour is set to 5.

Confusion Matrix

	Predicted Non-Fraudulent	Predicted Fraudulent
Actual Non-Fraudulent	TN 45007	FP 2
Actual Fraudulent	FN 8	TP 19

Table 3:KNN Confusion Matrix when N is 5

Scores

Accuracy - 0.999777955413447
 F1 - 0.7916666666666667
 Recall - 0.7037037037037037
 Precision - 0.9047619047619048

Changed the nearest neighbour value to 2, and the results for this were as follows:

Confusion Matrix

	Predicted Non-Fraudulent	Predicted Fraudulent
Actual Non-Fraudulent	TN 45002	FP 3
Actual Fraudulent	FN 13	TP 18

Table 4:KNN Confusion Matrix when N is 2

Scores

Accuracy --> 0.9996447286615152
 F1 --> 0.6923076923076923
 Recall --> 0.5806451612903226
 Precision --> 0.8571428571428571

This demonstrates that the KNN algorithm performed better when the value of N was larger, although the difference was comparatively negligible in terms of accuracy.

4.4 Working with Unbalanced Data with Logistic Regression

As demonstrated earlier, the nature of the data resulted in the Logistic Regression algorithm not detecting any fraudulent transactions, whereas KNN and Naive Bayes were able to detect the fraudulent transactions. This could be due to the nature of fraudulent transactions, which result in unbalanced data such as the data used in this research.

To deal with the unbalanced data, the methods that were recommended for handling unbalanced data were tested as demonstrated below:

a. Oversampling the Minority Class

Oversampling is adding more copies of the minority class. This process is a recommended choice when one lacks larger data samples to work with.

Oversampling the minority class resulted in a dataset with 337,570 fraudulent cases and 337,570 non-fraudulent cases.

The Logistic regression model was then trained using the oversampled data, and the results were as below:

1. The confusion Matrix was as below, and this demonstrated that the algorithm was capable of detecting fraudulent transactions in the oversampled data.

	Predicted Non-Fraudulent	Predicted Fraudulent
Actual Non-Fraudulent	TN 112,504	FP 4
Actual Fraudulent	FN 60	TP 20

Table 5: Oversampled data Confusion Matrix for Logistic Regression

2. The algorithm recorded an accuracy of 0.9994.

b. Undersampling the Majority Class

Under-sampling is the eliminations of some observations of the majority class. Undersampling is recommended for larger data samples such as those with millions of rows. However, a significant drawback to undersampling is the removal of valuable information.

Trained the Logistic regression model using the undersampled data, and the results were as below:

1. The confusion Matrix was as below, and this showed that the algorithm was capable of detecting fraudulent transactions in the undersampled data.

	Predicted Non-Fraudulent	Predicted Fraudulent
Actual Non-Fraudulent	TN 27,820	FP 84,688
Actual Fraudulent	FN 20	TP 60

Table 6: Undersampled data Confusion Matrix for Logistic Regression

2. The algorithm recorded an accuracy of 0.2476.

The undersampled data did not yield favourable results when the logistic regression algorithm was tested. A high number of false negatives were detected, which affected the algorithm's performance.

Based on the results of the tests done with the unbalanced data to improve the performance of the logistic regression algorithm, it is evident that oversampling the minority class yielded better results. Therefore, oversampling minority classes would be recommended for the unbalanced data in this research.

Data Attributes for Fraud Detection

As discussed in earlier chapters, the data contained 23 columns that were used for fraud detection. To test the attributes that were best suited for fraud detection,

1. Removed the scored data, SIM SWAP, AGE_SCORE and FIRST_TIME_SCORE banking transaction from the fraud detection system, and the results from running the algorithm were demonstrated in the table below.

	True Negative (TN)	False Positive (FP)	False Negative (FN)	True Positive (TP)
Logistic Regression	90018	0	53	0
Naïve Bayes	19	89997	0	55
K-nearest Neighbour	45002	3	8	23

Table 7:Confusion Matrix when scored data is dropped

2. Removed the static data that which included FORACID, reference number, Customer ID,

	True Negative(TN)	False Positive(FP)	False Negative(FN)	True Positive(TP)
Logistic Regression	90018	0	53	0
Naïve Bayes	90016	0	55	0
K-nearest Neighbour	45014	1	2	19

Table 8:Confusion Matrix when static data is dropped

3. Removed the transactional data, which included the following fields TRAN_DATE

TRAN_ID TRAN_AMT VALUE_DATE ENTRY_DATE
 PSTD_DATE VFD_DATE REF_NUM REF_AMT

	True Negative (TN)	False Positive (FP)	False Negative (FN)	True Positive (TP)
Logistic Regression	90013	0	58	0
Naïve Bayes	32390	12616	3	27
K-nearest Neighbour	45006	0	4	26

Table 9:Confusion Matrix when Transactional data is dropped

Below is a pie chart of the score based on the results obtained from using different data attributes. Based on these results, it is clear that the performance of KNN remains consistent and performs better than the rest of the algorithms. Additionally, it is noted that when the transactional data is removed KNN, performed marginally better than when the static and scoring data from the fraud detection system are removed.

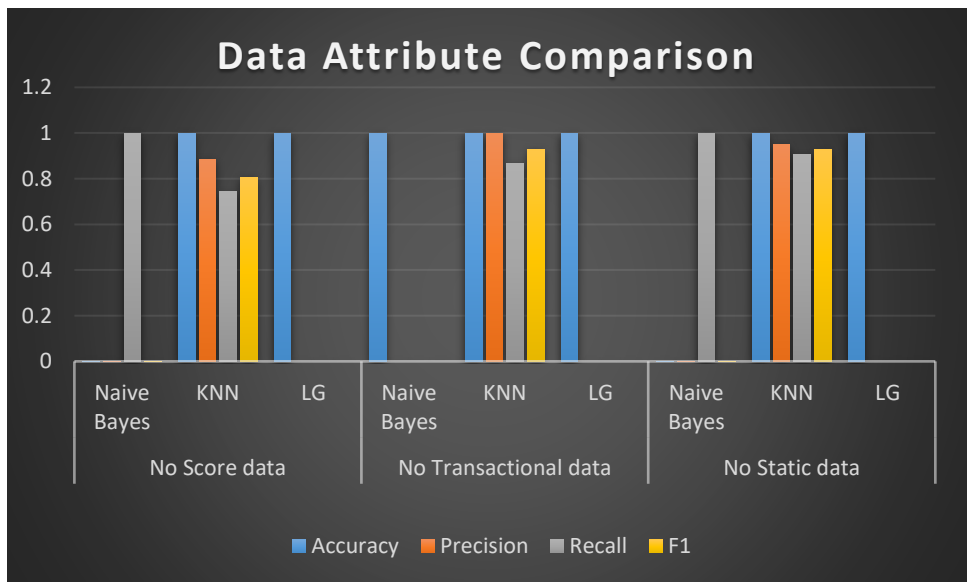


Figure 6:Data Attribute Performance Comparison

Achievement

The study's main objective was to analyse and evaluate various machine learning algorithms to determine their performance compared to rule-based systems. This was achieved by reviewing the performance of Logistic Regression, Naïve Bayes and KNN. According to observations KNN performed better than the other algorithms with a 99% accuracy level. This suggests that KNN can therefore be used in place of the rule-based systems, which require fraud analyst intervention to review each of the flagged transactions. Additionally, it was noted that once the data was oversampled and under-sampled for better performance for Logistic Regression, the algorithm did not perform as well as KNN which had the least number of false negatives(FN) and false positives(FP). In addition to the

performance of the different algorithms, based on the model's results being tested with different data attributes, it was determined that the algorithm's performance was better where the transactional data was removed, and the other data attributes remained. These attributes were CUST_ID, ACID FORACID, ACCT_NAME, SCHM_CODE, DELIVERY_CHANNEL_ID, PART_TRAN_TYPE, VFD_DATE, REF_NUM, REF_AMT, AGE_SCORE, FIRST_TIME_SCORE, SIM_SWAP_SCORE.

Conclusion

Fraudulent activities have increased over the COVID-19 period, and organisations with traditional fraud detection methods have proved to have underwhelming results because the human eye does not always capture anomalies in the banking system. Given the results obtained from this study, it is recommended that Equity Bank should use the KNN as its machine learning algorithm for fraud detection instead of other machine learning algorithms to replace the inaccurate rule-based fraud detection systems the bank uses.

Future Research

Future work should involve improving performance of the models by tuning parameters and hyperparameter. This process requires choosing hyperparameters for a learning algorithm and setting the value before the machine learning process commences. Adding more banking fraud transactions, with additional types of transactions labels, and experimenting with various sampling methods for class imbalance, should also be considered for future studies as they would make the fraud detection process more robust and reliable. Ultimately, using both supervised and unsupervised models for fraud detection with a limited number of labels and leveraging both methods should be considered to strengthen the models.

Bibliography

- Acfe.com. 2021. Association of Certified Fraud Examiners - Fraud 101. [online] Available at: <<https://www.acfe.com/fraud-101.aspx>> [Accessed 7 March 2021].
- Aghaei, E., 2017. *Machine Learning for Host-based Misuse and Anomaly Detection in UNIX Environment*. The University of Toledo.
- Ayyadevara, V., 2018. *Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R*. Apress.
- Banking and Financial Services. 2020. Retrieved 29 December 2019, from <https://similarity.com/banking-financial-services/>
- Bhatia, P., 2017. *Data mining and data warehousing*. Cambridge University Press.
- Brownlee, J., 2021. Logistic Regression for Machine Learning. [online] Machine Learning Mastery. Available at: <<https://machinelearningmastery.com/logistic-regression-for-machine-learning/>> [Accessed 5 July 2021].
- Buchanan, B. G., and Mitchell, T. M. 1999. Model directed learning of production rules. In Hayes-Roth, F., ed., *Pattern-directed inference systems*. New York: Academic Press.
- Card Fraud Losses reach \$27.85 Billion. 2019. *The Nilson Report*, (1164), Page 7.
- Chatfield, C. 1998. *The analysis of time series: An introduction (third edition)*. New York: Chapman and Hall
- Ciobanu, M., 2020. Why understanding your fraud false-positive rate is key to growing your business. [online] Thepaypers.com. Available at: <<https://thepaypers.com/thought-leader-insights/why-understanding-your-fraud-false-positive-rate-is-key-to-growing-your-business--1241130>> [Accessed 3 January 2021].
- COLVIN, G., 2020. The pandemic may be the greatest environment for business fraud in decades. [online] Fortune. Available at: <<https://fortune.com/2020/11/12/pandemic-corporate-fraud-scams/#:~:text=Fraud%20experts%20say%20every%20corporate,so%20they%20resort%20to%20trickery.>>> [Accessed 3 January 2021].
- Conversion-focused Fraud Detection | Bolt. 2020. Retrieved 29 December 2019, from <https://www.bolt.com/fraud/>
- Detelix Software Technologies Ltd. – Pay with Confidence. 2020. Retrieved 1 January 2020, from <http://detelix.com/>
- Enterprise Fraud Management | Clari5. 2020. Retrieved 10 January 2020, from <https://www.clari5.com/enterprise-fraud-management/>
- Ezawa, K., and Norton, S. 2005. Knowledge discovery in telecommunication services data using Bayesian network models. In Fayyad, U., and Uthurusamy, R., eds., *Proceedings of First International Conference on Knowledge Discovery and Data Mining, I00105*. Menlo Park, CA: AAAI Press.

Fraud Detection Algorithms | Fraud Detection using Machine Learning. 2020). Retrieved 1 January 2020, from <https://intellipaat.com/blog/fraud-detection-machine-learning-algorithms/>

Fraud Detection: How Machine Learning Systems Help Reveal Scams in Fintech, Healthcare, and eCommerce.2019. Retrieved 5 August 2019, from <https://www.altexsoft.com/whitepapers/fraud-detection-how-machine-learning-systems-help-reveal-scams-in-fintech-healthcare-and-ecommerce/>

Fraud Prevention | RSA Fraud & Risk Intelligence Suite. 2020. Retrieved 29 December 2019, from <https://www.rsa.com/en-us/products/fraud-prevention>

Graupe, D., 2016. *Deep Learning Neural Networks*. Singapore: World Scientific Publishing Company.

Harrison, O., 2021. Machine Learning Basics with the K-Nearest Neighbors Algorithm. [online] Medium. Available at: <<https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>> [Accessed 5 July 2021].

How to Choose Fraud Detection Software: Features, Characteristics, Key Providers. (2020). Retrieved 1 January 2020, from <https://www.altexsoft.com/blog/business/how-to-choose-fraud-detection-software-features-characteristics-key-providers>

Human Resource Management. 2019. Retrieved 5 August 2019, from <http://infosmarttech.com/machinelearningFraud.html>

Kapoor, A., 2019. *Hands-On Artificial Intelligence for IoT: Expert machine learning and deep learning techniques for developing smarter IoT systems*. Birmingham: Packt Publishing Ltd.

Kount Complete Product for Digital Fraud Prevention | Kount. 2020. Retrieved 9 January 2020, from <https://www.kount.com/fraud-detection-software/kount-complete>

Littman, A., 2011. *The Fraud Triangle: Fraudulent Executives, Complicit Auditors and Intolerable Public Injury*. CreateSpace Independent Publishing Platform.

Maruti Techlabs. 2021. How Machine Learning Facilitates Fraud Detection?. [online] Available at: <https://marutitech.com/machine-learning-fraud-detection/#Benefits_of_Fraud_Detection_via_Machine_Learning> [Accessed 7 March 2021].

Moon, W., & Kim, S. 2017. Adaptive Fraud Detection Framework for FinTech Based on Machine Learning. *Advanced Science Letters*, 23(10), 10167-10171. doi: 10.1166/asl.2017.10412

Omnisci.com. 2021. What is Fraud Detection and Prevention? Definition and FAQs | OmniSci. [online] Available at: <<https://www.omnisci.com/technical-glossary/fraud-detection-and-prevention>> [Accessed 9 March 2021].

Oniyilo, T., 2016. *Payroll fraud detection and prevention audit expert system*. [Place of publication not identified]: Lulu Com.

Pascual, A., Marchini, K. and Miller, S., 2017. *2017 Identity Fraud: Securing the Connected Life*. [online] Javelin. Available at: <<https://www.javelinstrategy.com/coverage-area/2017-identity-fraud-securing-connected-life>> [Accessed 30 July 2021].

Pun, J. and Lawryshyn, Y. 2012. Improving Credit Card Fraud Detection using a Meta-Classification Strategy. *International Journal of Computer Applications*, 56(10), pp.41-46.

- Rajeshkumar, I. (2019). Fraud Detection in Banking Industry and Significance of Machine Learning. Retrieved 5 January 2020, from <https://medium.com/engineered-publicis-sapient/fraud-detection-in-banking-industry-and-significance-of-machine-learning-dfd31891a0b4>
- Rouse, M. (2019). What is fraud detection?. Retrieved 5 January 2020, from <https://searchsecurity.techtarget.com/definition/fraud-detection>
- Sando, S., 2021. *Consumer Preference Drives Shift in Authentication*. [online] Javelin. Available at: <<https://www.javelinstrategy.com/coverage-area/consumer-preference-drives-shift-authentication>> [Accessed 30 July 2021].
- SAS Fraud Management. 2020. Retrieved 29 December 2019, from https://www.sas.com/en_us/software/fraud-management.html
- Team, C. (2015). Bank Fraud - Definition, Examples, Cases, Processes. Retrieved 5 January 2020, from <https://legaldictionary.net/bank-fraud/>