



E-COMMERCE FRAUD DETECTION USING SUPERVISED AND UNSUPERVISED MACHINE LEARNING ALGORITHMS.

LOVE J. AGBANIMU
21120404

Supervisor
(Dr. AbdulRahman Alsewari)

September 2022

A dissertation to be submitted in partial fulfilment of the requirements
for the degree of Master of Science in Big Data Analytics

School of Computing, Engineering and the Built Environment
Birmingham City University

AKNOWLEDGEMENT

Firstly, I thank God for the grace, wisdom and understanding of this work.

My sincere gratitude goes to my supervisors Dr. AbdulRahman Alwaseri and Sara Hassan for their continuous support, patience and guidance for this master's dissertation which has been of great impact in the progress and completion of my project work.

Lastly, I am indebted to all my friends and my family: to my father, you have been of great support through this period, my mother, your prayers and love will forever be cherished, my siblings, nephews, nieces and in-laws, I love you all.

ABSTRACT

The reliability and performance of real time fraud detection techniques has been a major concern for the e-commerce institutions as traditional fraud detection models couldn't cope with the emerging dynamic, diverse and innovative fraud patterns that deceive their platforms. Apart from the concept drift experienced by the fraud detection systems where fraudsters try to circumvent the system by modifying their attack after the decision system was successful in blocking their previous fraud attempt, genuine users can also be subject to concept drift which makes the current fraud detection systems to be less accurate in stopping fraud.

This study looks at the holistic view of fraud detection in e-commerce platforms and proposes a fraud detection framework that can detect anomalous transactions quickly and accurately, as well as maintaining the efficiency with minimum input from subject matter experts. Firstly, this study will focuses on Data Cleaning, Feature Engineering and Feature Selection, alongside the raw attributes of the transactional data, novel features for e-commerce fraud detection are created and evaluated. The second focus will be on modeling, which will be performed on a real-life transactions dataset, provided by a large e-commerce institution. The results of the modeling will bring about new models which when provided a transactional dataset will detect fraudulent transactions with acceptable accuracy and reduced False Negative Rate.

Keywords: Feature Engineering, Feature Selection, Supervised and Unsupervised Machine Learning Algorithm, E-commerce platforms, Fraud Detection models, Expert Systems approach, Deep Learning approach, XGBoost model, LGBM model, Random Forest model.

TABLE OF CONTENTS

1.	Introduction	1
1.1.	Background of Study	1
1.2.	Problem Statement	3
1.3.	Research Aims and Objectives	4
1.4.	Scope of Study	5
1.5.	Research Questions	5
1.6.	Significance of Study	6
1.7.	Project Structure	6
2.	Literature Review	8
2.1.	Fraud in E-commerce Platforms	8
2.2.	Rule-Based Expert Systems Approach to Fraud Detection	9
2.3.	Un-supervised Machine Learning Approach	11
2.4.	Supervised Machine Learning Approach	13
2.5.	Deep Learning Approach	16
2.6.	Summary and Conclusion	18
3.	Methodology	19
3.1.	Background	19
3.2.	Research Approach	20
3.3.	Research Framework	20
3.4.	Data Collection	23
3.4.1.	Historical Data Collection	23
3.4.2.	Synthetic Data Generation	25
3.5.	Feature Engineering	26
3.6.	Model Description	26
3.6.1.	Clustering Algorithm	26
3.6.2.	Bagging Models	27
3.6.3.	Boosting Models	27
3.7.	Performance Evaluation Metrics	28

3.8.	Baseline Model	30
3.9.	Conclusion	31
3.10.	Summary of Procedure	31
4.	Data Analysis and Model Implementation	32
4.1.	Model Development	32
4.2.	Data Preprocessing	33
4.2.1.	Data Description	33
4.2.2.	Exploratory Data Analysis	36
4.3.	Feature Engineering	43.
4.4.	Model Training and Implementation	46
4.4.1.	Random Forest Model	46
4.4.2.	Gradient Boosting Models	47
5.	Evaluation and Validation	50
5.1.	Modelling Result and Analysis	50
5.1.1.	XGBoost Model Performance	51.
5.1.2.	LGBM Model Performance	52
5.1.3.	Random Forest Model Performance	53
5.1.4.	Performance Comparison	
5.2.	Model Robustness	55
5.3.	Conclusion	56
6.	Conclusion and Recommendation	57
6.1.	Overview	57
6.2.	Research Contributions	57
6.3.	Recommendations for Future Work	58

LIST OF TABLES

1. Table 3.1: Confusion Matrix	28
2. Table 4.1: Transaction Dataset Features	34
3. Table 4.2: Distributions of the fraudulent and non-fraudulent transactions	36
4. Table 4.3: Number of Fraudulent and Non-Fraudulent Transactions by Amount Range	39
5. Table 5.1: Datasets split into training and validation	50
6. Table 5.2: Models Performance	54
7. Table 5.3: Models Robustness	55

LIST OF FIGURES

1.	Figure 3.1: Research Framework	21
2.	Figure 3.2: Research Framework Sections	22
3.	Figure 3.3: Barchart showing the fraud and non-fraud counts	24
4.	Figure 3.4: SVM Model Evaluation Report	30
5.	Figure 4.1: Dataset snapshot	33
6.	Figure 4.2: Distributions of fraudulent and non-fraudulent transactions	36
7.	Figure 4.3: Distributions of Non-fraudulent transactions	37
8.	Figure 4.4: Distributions of Fraudulent transactions	38
9.	Figure 4.5: Number of Fraudulent and Non-Fraudulent Transactions by Amount Range	38
10.	Figure 4.6: Number of Fraudulent and Non-Fraudulent Transactions by hour ..	39
11.	Figure 4.7: Number of Fraudulent and Non-Fraudulent Transactions by Day of week.....	40
12.	Figure 4.8: Transactions Amount by Year and Month.....	40
13.	Figure 4.9: Transactions Amount by Holiday and Month.....	41
14.	Figure 4.10: Transactions Amount Distribution by Week Days.....	42.
15.	Figure 4.11: Number of Fraudulent and Non-Fraudulent Transactions by clustering.....	44
16.	Figure 4.12: Distributions of the fraudulent and non-fraudulent transactions after sampling.....	45
17.	Figure 4.13: Flow of Random Forest Algorithm Operations.....	46
18.	Fig 4.14: Gradient Boosting Algorithm Operations.....	48
19.	Figure 4.15: Fraud Detection Modeling Process.....	49
20.	Figure 5.1: XGBoost Model Evaluation Report	51
21.	Figure 5.2: LGBM Model Evaluation Report	52
22.	Figure 5.3: Random Forest Model Evaluation Report	53

CHAPTER ONE INTRODUCTION

1.1 Background to Study

Recent advances in technology have led to an increase in fraud, which costs billions of dollars annually worldwide (Chang, et al. 2017). Examples include online banking fraud (Wei, et al. 2013), credit card fraud transactions in e-commerce (Sahin, et. al, 2013), and telecommunication fraud. E-commerce has become an important part of the global business sector. Following the development of the internet, the retail environment saw a significant transformation like that of many other businesses. As a result of the continuous digital transformation of modern life, consumers in all countries today benefit from online purchases. The number of digital shoppers is growing each year as global internet access and adoption rapidly increase. For instance, according to a report by Weng et al. (2018), Taobao (part of the Alibaba Group) had 443 million active consumers in the year 2016, and eBay had above 164 million active customers in 2016, and Amazon had 310 million active customers in 2016.

E-commerce today effectively links consumers with producers, retailers, and independent merchants, giving them access to a convenient, quick, and trustworthy method of shopping. Online shopping is becoming increasingly popular over traditional means because of its many benefits. Manufacturing facilities, independent retailers, and ecommerce service providers have benefited economically from the rapidly growing ecommerce retail sales. For instance, according to reports, the Gross Merchandise Volume (GMV) of Taobao reached US \$320 billion in the 2017 fiscal year, the GMV of Amazon reached US \$149 billion in 2016, and the GMV of Jingdong reached US \$101 billion in 2016 (Weng et al., 2019).

Global e-commerce and online customer behaviour were significantly impacted by the coronavirus (COVID-19) pandemic. It is observed that people shop on average around 10-30% more online. With other e-commerce businesses seeing an approximately 50% increase, grocery e-commerce saw a 250% increase which eventually dropped to as

much as 150% above the original level (Paintal, 2021). This enormous volume of e-commerce transactions increases the risk of new issues, including e-commerce transaction fraud.

Retail merchants now have a plethora of new prospects thanks to the tectonic shift that e-commerce has brought about in the retail sector. Although technology makes doing business with other businesses and customers more convenient, it has unluckily also made them vulnerable to significant threats from skilled fraudsters who engage in various forms of online transaction fraud and online service abuse. As online sales increase, fraudsters are more able to target merchants who are new to e-commerce, inexperienced with it, or lacking in the means to implement advanced security measures. For all online retailers worldwide, preventing the loss brought on by unplanned fraud assaults without reducing the revenue from legal transactions has always been a critical challenge. This demand has led to various fraud detection systems that leverage techniques, including machine learning, to find optimal solutions to these threats.

Because fraud trends are so varied and dynamic, preventing fraudulent e-commerce transactions is a significant difficulty. Fraudsters frequently alter their attack vectors when they either discover a new vulnerability in the fraud protection system to exploit or feel that the merchants successfully stop their nefarious activities. To avoid being detected by e-commerce fraud protection systems, fraudsters often go to great lengths to act like legitimate consumers. It is difficult for online merchants to effectively distinguish between fraudsters and genuine customers because, like fraudsters, genuine customers' online transaction behaviour changes with time. The ability of the study into e-commerce fraud identification to give consistent management of the platforms has been constrained, in contrast to the development in techniques of the fraudsters.

Given the constant regular activity on e-commerce platforms, it could be challenging to discover anomalies that indicate fraud has happened using conventional methods. To successfully defend against the prevalence of fraud in a constantly changing threat landscape, organizations must solidify their platform's security infrastructure and make

use of technologies like machine learning to ensure proactive security control. It is crucial to promptly and aggressively react to frauds that would otherwise blend in with regular transactional operations, as well as automatically identify frauds that would otherwise go undetected.

Along with the opportunities of 2020 poured over into 2021 is the continued growth of global e-commerce sales. However, the increase in e-commerce purchases and first-time online shoppers has increased online shopping fraud. E-commerce merchants must comprehend what they are up against because of this growing threat, the evolution of conventional fraud schemes, and the introduction of new tactics.

Machine Learning is a critical component of e-commerce fraud detection, as it provides the foundation for the solution's efficiency. Machine learning is used alongside deep learning to identify, categorize, and detect fraud in an e-commerce platform. Machine learning models are constructed to handle complicated problems and increase the effectiveness of fraud detection in e-commerce platforms because of the heterogeneity in the explosion of data generation. This research focuses on the possibility of creating a new real-time fraud detection framework to detect fraud in electronic commerce transactions. The intention is to use advanced machine learning techniques to improve the fraud detection rate in e-commerce transactions.

1.2 Problem Statement

Fraud detection is critical to effective retail business performance by helping to monitor consumer transactions to identify anomalies, maximize performance for a good customer experience, and protect against fraudulent attacks. Numerous researchers have investigated methods and techniques to control fraud on e-commerce platforms.

However, these studies have focused on different approaches to the problem and have overlooked several essential features of machine learning algorithms that can produce effective results—as a result, limiting the knowledge of the area of focus.

This research will therefore evaluate different machine learning ensemble algorithms like

Random Forest, Gradient Boosting, Extreme Gradient Boosting, and others to compare their limitations and strengths and increase their effectiveness in correctly classifying a transaction as fraud or legitimate in an online retail context for e-commerce settings.

1.3 The Research Aims and Objectives

The study aims at creating a fraud detection model that will aid in the advancement of the e-commerce industry by building a framework that employs machine learning techniques on e-commerce transaction data to reduce false positives and enhance detection rates. This research will create a fraud framework that can be trusted by utilizing supervised and unsupervised machine learning techniques. By combining these strategies, the fraud detection rate can be improved. The quantity and accuracy with which a fraud detection system can identify fraudulent occurrences are indicators of its effectiveness. The specific goals of this research are;

1. To examine the present status of research on e-commerce platform fraud detection and to identify significant problems, current strategies, and practical ways to enhance a fraud detection system's performance.
2. To investigate the availability of data across multiple e-commerce channels to create the appropriate datasets and criteria for analysis.
3. To explore different methods of machine learning for analyzing real-time multidimensional e-commerce data.
4. To implement an e-commerce fraud framework that blends supervised and unsupervised machine learning models on a suitable platform for fraud detection.
5. To evaluate the effectiveness of the proposed fraud detection model, as well as the efficiency of the approaches that were used.
6. To offer pertinent findings, suggestions, and potential future research directions.

1.4 Scope of Study

This study aims at using machine learning algorithms to assess and analyze e-commerce fraud. The research analyzes e-commerce fraud to create a detection system that

effectively detects fraudulent attacks. To enhance the effectiveness of these models and suggest the best fraud detection system, an unsupervised machine learning algorithm and various ensemble supervised machine learning algorithms, ranging from random forest classifiers, gradient boosting algorithms, extreme gradient boosting algorithms, and others, will be implemented to determine their strengths and limitations.

1.5 Research Questions

1. What is the status of research on e-commerce platform fraud detection, and what options are there for enhancing the effectiveness of a fraud detection system?
2. What data will be appropriate for addressing fraud detection on e-commerce platforms?
3. What are the contributions of new technologies such as machine learning in ecommerce fraud detection systems?
4. How can an e-commerce fraud framework that blends supervised and unsupervised machine learning models be implemented?
5. How effective are the proposed fraud detection models?

1.6 Significance of Study

The results of this study will impact various managerial tasks in e-commerce firms by helping to build ways to track online transaction activity to spot abnormalities, boost efficiency, and safeguard the platform against attacks.

By highlighting and assessing the application of several essential components of machine learning techniques in fraud detection tactics and approaches for the industries that are susceptible to fraud assaults, this research will add to the body of knowledge on ecommerce fraud detection. This study will also benefit e-commerce organizations by addressing the present gap in knowledge regarding the use of machine learning techniques for e-commerce fraud detection and providing value.

1.7 Project Structure

In Chapter 1, the research problem has been introduced. The Research scope, questions, and objectives have been identified, and the significance of the research has been stated. In Chapter 2, various research papers will be examined and reviewed in the context of fraud detection. The nature of current research, methodologies and suggested solutions to issues directly connected to fraud detection are investigated.

In Chapter 3, the theoretical research framework will be described. The use of a quantitative research strategy will be defended, and a more comprehensive design of the study will be covered.

In Chapter 4, the system framework architecture and details on how the selected supervised and unsupervised machine learning models are implemented will be described. Real e-commerce transactional data will be investigated, and a detailed exploratory analysis will be provided.

In Chapter 5, thorough methods will be used to evaluate and contrast the model performances using well-defined indicators and datasets. This chapter will as well describe the hyperparameters utilized by the implemented models.

In Chapter 6, the thesis conclusion, outlines of what has been accomplished, a summary of the study, and suggestions for future work will be presented.

CHAPTER TWO LITERATURE REVIEW

The studies that have been conducted on fraud detection and associated works are covered in this chapter. To review the research literature, the literature was first collected depending on the goals and issues of the study. Different journals and publications databases were searched using keywords like E-commerce Services, Fraud Detection System, Machine Learning (ML), Intrusion Detection, Ensembling Models, Artificial Neural Network (ANN), and others. By evaluating an author's response to the research question, key ideas, theories, methodologies, and models employed, articles and papers that were significant to the subject of this thesis were chosen from the relevant literature. The results and review of the outcomes came next. Considering this, it will examine the accuracy and constraints of various methodologies and approaches while taking this into account, the definitions of concepts and problems that impact the detection of fraudulent actions. The chapter begins with an introduction to e-commerce sector fraud and the most recent rising risks. The study of fraud detection methods, including conventional rule based expert systems, deep learning, unsupervised and supervised machine learning methods, will come next. The chapter will end with a description of the course this research will take.

2.1 Fraud in E-commerce Platforms

Fraud is described by the Association of Certified Fraud Examiners (ACFE) as "any activity that relies on deceit in order to acquire a gain." When a person "knowingly misrepresents the truth or conceals a material fact to encourage another to act to his or her detriment," fraud becomes a crime (Black's Law Dictionary). Fraudulent actions have taken place in technical systems in various aspects of daily life, including mobile communications, telecommunication networks, network traffic, E-commerce, and online banking. People, the government, and businesses all suffer significant financial losses due to these crimes. As a result, fraud detection has grown in importance as a topic for investigation.

According to a Global Fraud Report released by Cybersource and the Merchant Risk

Council (MRC) in 2021, it has been demonstrated that during the previous two years, COVID-19 inception and the resulting restrictions on traditional, offline trade both accelerated online sales and increased the prominence of e-commerce as a vital sales channel for many merchants globally. As a result, nine out of ten merchants regarded preventing e-commerce fraud as very important to their entire company strategy, which led to a five-fold rise in expenditure on fraud as a percentage of e-commerce sales since 2019. Additionally, it was found that although the number of fraud assaults increased, the variety of fraud attacks that merchants encountered decreased, with identity theft, phishing, card testing, and friendly fraud being the most common attack types affecting the highest proportions of retailers globally.

The trends above and figures are rough estimates, but they do highlight the significance of having a system in place for fraud detection and prevention that is efficient and successful. The various research components of the fraud detection system are described in the subsequent sections.

2.2 Rule-Based Expert Approach to Fraud Detection

Rule-based expert systems that utilize rules are traditionally employed to identify fraudulent transactions. These systems are predicated on the subject matter expert's knowledge and intuition. When a customer reports an unrecognized transaction that appears on their bank statement, for example, this typically entails a detailed manual analysis of the questionable situation. Investigations conducted later may reveal recently developed fraud techniques employed by the fraudster. Then, regulations are created or changed to reduce future fraud exposure.

"Thresholds" are chosen by fraud subject matter experts in most expert-based fraud detection systems. Such rules are challenging to administer, uphold, and put into practice. The regular regulation updates and evolving fraud schemes and strategies also demand a complicated ecosystem. One drawback of a rule-based system is that fraudsters can learn the thresholds and the rules through trial and error to come up with creative workarounds quickly. Another drawback of the rule-based approach is that it is difficult to

generalize to newly developing fraud patterns because the rules are built based on previously known fraud incidents (Van Vlasselaer et al., 2015). A decision must be taken regarding when to update, remove, or add new thresholds and rules to the existing ones to allow this system to adapt to newly established patterns.

Weng et al. (2018) created an effective and scalable Anti-Fraud system (ATF) to identify e-commerce frauds for large-scale e-commerce platforms, and they concurrently implemented it on the Open Data Processing Service large-scale computing platform (ODPS). The ATF system consists of three components: Preprocessor, Graph-Based Detection Module (GBD), and Time Series Based Detection module (TSD). The preprocessor is used to process raw data and get the data ready for the GBD and TSD modules. The GBD, on the other hand, uses a propagation method to award each item a fraud score based on the user-item bipartite graph and a small sample of verified dishonest users. The fraud items are then selected based on their fraud scores. TSD was based on the observation that the traffic time series of a new fraud item is likely to display differently from other traffic time series when a new fraud pattern or item becomes a fraud item. The authors used two actual large-scale e-commerce datasets to evaluate the ATF. Their evaluation's findings show that ATF can attain precision and recall rates of 0.97+, indicating that it is highly efficient. More significantly, the authors implemented the ATF on Alibaba's Taobao platform, one of the biggest e-commerce platforms in the world. The evaluation's findings revealed that ATF could obtain a 98.16 percent accuracy rate on Taobao, which again indicates that ATF is very effective and deployable in practice.

Jha et al. (2012) implemented a transaction aggregation technique to identify credit card fraud. To estimate a model and identify fraudulent transactions, the authors aggregated transactions to record customer purchasing behaviour prior to each transaction. For transaction aggregation and model estimation, the authors used actual credit card transaction data from a global credit card operation. As a result of the model accurately classifying transactions using derived features, the authors concluded that transaction aggregation is a valuable method for detecting fraud.

The study in this section showed that rule-based expert fraud detection systems could detect fraud effectively, although labour-intensive and requiring manual input from subject

matter experts. One of the primary issues with the research is its ability to generalize a novel fraud pattern that has not been observed previously. There is a need for a more efficient fraud detection system that requires less human skill and provides a quicker feedback loop when anomalies in transactions are discovered.

2.3 Unsupervised Machine Learning Approach.

There are numerous unsupervised techniques available to identify fraudulent transactions in a range of industries, including telecommunications, remote banking, and credit cards. Unsupervised approaches look for the customers' accounts, transactions, and other features that deviate the greatest from the norm. The norm is defined as the average behaviour of a customer or the typical customer behaviour over various periods. (Baesens et al., 2015). Transactions that deviate from the norm are flagged as "fraud" to draw attention to them. Unsupervised techniques commonly referred to as outlier identification, attempt to model the distribution of the expected behaviour from the perspective of fraud detection and highlight any observation that deviates the most from this normal behaviour. This technique works by grouping together comparable observations, which does not require previously labelled data.

Clustering algorithms are one of the significant subsets of unsupervised learning. The clustering method seeks to identify clusters of related observations within the data that are characterized by the highest degree of similarity within a cluster and the highest degree of dissimilarity across clusters (Sinaga and Yang, 2020). Clustering can be used in the retail industry to identify both regular transactions and fraudulent transactions. Customers' sociodemographic, behavioural, or other account characteristics, like a customer transaction, can be used to derive the different clusters in the dataset.

Rashmi et al. (2018) employed Self-Organizing Maps (SOM) visualization techniques to identify fraudulent accounts. To visualize accounts in 2D space for simple analysis by a human expert and to combine it with classification algorithms to increase the detection rate, the author used SOM. SOM's U-Matrix is used to build the account clusters by measuring the average distance between the neuron and its neighbours. This suggested

method is applied to intrusion detection systems, credit card fraud, and telecommunications fraud.

To identifying credit card fraud, two unsupervised machine learning techniques based on the AE model and the restricted Boltzmann machine (RBM) model are proposed in (Pumsirirat, 2018). Both AE and RBM perform well in terms of fraud detection, as illustrated in study. However, it is demonstrated that AE performs better than RBM. Zamini and Montazer, (2018) implemented an unsupervised AE-based clustering technique for the purpose of identifying credit card fraud. By selecting an acceptable threshold of AE reconstruct, the AE-base clustering technique was able to achieve a good classification performance.

Singh and Narayan (2012) implemented a Hidden Markov Model for credit card fraud detection. The author used HMM to profile client spending patterns on credit card transactions by categorizing incoming transactions into different categories, such as low, medium, or high, and then flagging a transaction as fraudulent if they deviate from a predetermined threshold value. The findings demonstrated that HMM effectively identified outliers with a significant false-positive rate. Bhati and Sharma (2015) combined K-means and Hidden Markov Model to detect credit card number fraud. The authors combined these two models using K-means for clustering and utilizing HMM to identify outliers within the clusters and then implemented Luhn Algorithm to validate the results of the models. The paper's conclusion emphasizes that using HMM enhances results when compared to using K-Means alone.

Chougule et al. (2015) explain the combination of genetic algorithms and K-means to detect credit card fraud. They establish three different clusters: low, medium, and high risk, in which the transaction will be included in any of them. A genetic algorithm was used to increase efficiency when stable centroids were attained. This method reduces the requirement for prior knowledge of fraud patterns and allows for the incorporation of current knowledge into a semi-supervised clustering methodology.

This study shows that the clustering methodology is the most effective unsupervised machine learning method for finding outliers. This approach to fraud detection is practical

and frequently used in ensembled approaches because of the non-parametric nature of forming clusters and subclusters. This method reduces the requirement for prior knowledge of fraud patterns and allows for the incorporation of current knowledge into a semi-supervised clustering methodology.

2.4 Supervised Machine Learning Approach.

There have been numerous research looking at how to improve the performance of different models on fraud detection since the introduction of machine learning and deep learning methods to the big picture. Methods for detecting fraud may be described as either supervised or unsupervised techniques. In supervised approaches, samples of both fraudulent and legitimate data are used to build models, which are later used to classify new observations as either fraudulent or legitimate. Additionally, it can only be used to find frauds of the same kind that have already happened. Some supervised machine learning methods include Random Forest, Logistic Regression, Random Forest, Decision Trees, Support Vector Machines, Artificial Neural networks, Gradient Boosting, Extreme Gradient Boosting, and so on.

Support vector machine (SVM), naive Bayes (NB), feed-forward neural network (NN), and nine other machine learning models for credit card fraud detection are examined in (Randhawa et al. 2018) Additionally, to improve performance, two ensemble learning mechanisms—adaptive boosting (AdaBoost) and majority voting (MV)—are integrated with the twelve models. SVM paired with AdaBoost (referred to as SVM + AdaBoost) and NN and NB combined with MV (referred to as NN + NB + MV) exhibit comparable high performance, according to thorough performance comparisons.

Nanduri et al. (2020) offer two strategies, fraud islands and multi-layer models, to improve the machine learning model's capacity for detecting e-commerce fraud. Link graph aggregated features that can more effectively give some vital information about the concealed fraud patterns were developed using fraud island. Using the multi-layer modeling technique, the authors developed three sub-models for transactions assigned fraud labels by various risk prevention systems. It is thought that more fraud in various

types of fraud patterns could be discovered by employing the fraud labels established by various internal and external risk systems. To demonstrate the effectiveness of the multilayer model, the authors created two datasets for each of the three portfolios used in the research, one for in-time evaluation and the other for out-of-time evaluation. The authors then compared the performance of the Multi-layer (ML) models and Long Term models that did not utilize the outputs from other models and showed that the Multi-layer (ML) model outperforms the Long Term Model across the three different portfolios. The authors gathered and used seven months' worth of actual e-commerce transaction data, which comprised the encrypted customer information and fraud label tagged, to show how fraud islands can improve machine learning model performance. The investigation demonstrates that, with a 3% improvement in the area under the curve (AUC) score, the Gradient Boosting Tree model with aggregate features from link analysis outperformed the model without them.

Dornadula and Geetha (2019) developed a method to detect credit card fraud, in which customers are grouped according to their transactions, and then behaviour patterns are extracted to develop a profile for each cardholder. The authors established three assessment scores by applying different supervised learning classifiers to three groups. The dynamic parameter changes cause the system to adapt in time to the transaction behaviours of new cardholders and enable fraudulent activities to be caught to a greater extent.

By combining the bagging strategy, which can lower the variance of the classification model by resampling the data, and the boosting technique, which lowers the model's bias, Bian et al. (2016) devised an ensemble approach for financial fraud detection. The data are split into two classes using the bagging procedure, then fitted into the boosting classifier. The distribution of the minority and majority classes was balanced via oversampling and undersampling for each sub-training data set. The sub-classifier voting combination was then used to generate the outcome. This method's flaw is that it performs worse when the classifiers are combined.

Randhawa, K., (2018) developed a hybrid technique that combines AdaBoost and majority voting techniques and twelve standard models to improve the accuracy of credit card fraud detection. Naive Bayes (NB), k-nearest neighbour (k-NN), and logistic regression (LR), three techniques for detecting credit card fraud, are proposed in (Awoyemi, J., 2017). The k-NN approach with a $k = 3$ achieves the best classification outcome. The paper used a random data resampling approach to overcome the issue of data imbalance which enhanced the performance of the k-NN approach.

Using data mining techniques, Agarwal and Mitta (2012) suggested a hybrid method for identifying fraud attacks that involve anomalous network traffic. The entropy of network features and Support Vector Machine (SVM), two anomaly-based detection techniques that the authors worked with, were combined into a hybrid model for improved intrusion detection. For this investigation, traffic data provided by MIT Lincoln Laboratory was used. The entropy-based intrusion detection system measures the degree of randomness of a few chosen normalized network parameters, and any divergence from a specified range denotes abnormal network traffic. Support Vector Machine was trained on several network features to produce a model that distinguishes between regular traffic and attack traffic. The entropy method can represent the network, and the support vector machine has the advantage of reasonable classification. The authors suggested a hybrid approach that takes advantage of both methodologies' advantages to increase the models' accuracy. By calculating the normalized entropy of network features and sending it to an SVM model for learning, the hybrid approach then categorizes network traffic as either legitimate or malicious. The experimental outcome of this study demonstrates that, by decreasing the rate of false alarms and boosting the identification of anomalies, the hybrid method, compared to the individual strategies, proves to be the best model.

2.5 Deep Learning Approach.

Deep learning models are built based on artificial neural networks, which handle a variety of practical pattern recognition and classification issues. The inspiration for neural networks came from the parallel computation seen in human brains, which allowed the

neural connections to be adjusted accordingly (Rojas, 2013). Deep learning provides many techniques for learning several layers of representation that uncover undiscovered connections and patterns in the data and provide a powerful generalization.

Guo et al. (2019) offered two innovative attack strategies to address the discrete optimization brought on by the deep fraud detector's susceptibility to minute changes in input transactions. The first strategy uses enhanced iterative search (AIS), which repeatedly searches for "simple" yet strong perturbations. The second method is referred to as the Rounded Relaxation with Reparameterization (R3), and it rounds the result of applying reparameterization techniques to a simple and unrestricted optimization problem. Using millions of actual transactions from Taobao e-commerce platforms, one of the biggest in the world, they carried out a thorough experimental evaluation of the fraud detector that had been installed. Their findings demonstrate that the deployed model is highly vulnerable to fraudulent attacks due to the drop in average precision. In Addition, the model they developed via an adversarial training approach was more substantial and robust against attacks and outperformed the adaptations in terms of performance.

Chouiekh and Haj (2018) compared the ability of convolution neural networks to predict fraudulent occurrences in mobile communication networks to other more established machine learning algorithms. With an accuracy of 82%, their results show that the Deep Convolutional Neural Network architecture outperformed the Support Vector Machine, Random Forest, and Gradient Boosting Classifier algorithms and produced the best results.

Fu et al. (2016) represented credit card transaction data as a feature matrix using a convolutional network for fraud detection to identify the inherent patterns in the data. The authors used cost-based sampling to solve the issue of class imbalance in the dataset. Their proposal measured customer preferences over various time frames and calculated trade entropy using model customer profiles. Latent variables were generated and structured in a way that can be easily fed through the CNN's filter bank as part of the feature engineering process. They ran their model on actual banking data, and empirical evidence showed that it outperformed the model used by the current bank. Chen and Lai

(2021) applied a Deep Convolutional Neural Network (DCNN) based financial fraud detection scheme using a deep learning algorithm on a random sample of 5 million transactions collected over 24 hours. The result of the authors modeling and analysis shows that when a large volume of data is involved, the fraud detection accuracy can be enhanced by using this technique.

Nathan et al. (2018) in their study, address the problem of unsupervised feature learning using nonsymmetric deep autoencoders (NDAE) and provide a unique deep learning method for intrusion detection. The model has so far produced encouraging findings that show advances over current methods and a great potential for usage in contemporary Network Intrusion Detection Systems. To perform network traffic classification, Hyun-Kyo et al. (2019) built five deep learning models utilizing the convolutional neural network (CNN) and residual network (ResNet). After comparing the CNN and ResNet deep learning models' f1 scores for network traffic categorization performance on packetbased datasets, the authors concluded that ResNet outperforms the CNN model.

Wang et al. (2017) developed a deep-learning-based system for detecting transaction fraud, which was then implemented on one of the biggest e-commerce sites in China with over 220 million active users, JD.com. The model uses recurrent neural networks to simulate click sequences and neural network-based embedding to capture comprehensive information on user click behaviours. The fraud detection system also optimizes application-specific designs, such as unbalanced learning, real-time detection, and incremental model update. The authors demonstrate that the fraud detection system achieves over three times improvement over the current fraud detection methodologies by using production data for more than eight months.

Batani (2017) investigated a real-time adaptive system for detecting fraud in credit cards. An artificial neural network, a hidden Markov model, and a one-time password are used in the solution. The user's profile is collected from the bank database by the system following the One-Time password's successful authentication, and this profile is then categorized using Artificial Neural Networks. The user's financial profile was subsequently produced using a Hidden Markov Model. For fraud detection, a mix of the three techniques

was applied. The suggested method was evaluated using simulated data, and fraud was effectively identified and stopped.

2.6 Summary and Conclusion.

The research-reviewed literature outlines the overall issue and provides a few remedies, each of which has employed a unique strategy to address it. This chapter has shown that several studies focus on fraud detection. The variety of approaches, including unsupervised, supervised, and deep learning strategies, to detecting fraud are covered in the papers. According to the research, traditional methods of fraud detection based on domain knowledge, expert intuition, and experience could not keep up with developing fraud, necessitating the employment of other methods. Most of the research on detecting fraud has been conducted using popular machine learning algorithms like support vector machines, decision trees, hidden Markov models, and others. The use of ensemble machine learning techniques to identify fraudulent events in e-commerce intuitions has not received much investigation. Further research is necessary to see whether the detection rate of fraud can be increased by combining these ensembling techniques with unsupervised learning techniques. Considering all the findings from the literature review and the author's experience, the parts of the research mentioned in the proposal to use sophisticated machine learning algorithms to improve fraud detection remain valid.

CHAPTER THREE METHODOLOGY

3.1 Background

This situation involves selecting the most appropriate method of several available alternative methods of detecting fraudulent activities in retail e-commerce platforms. As discussed in the previous chapters, e-commerce services face increased fraudulent activities across their platforms. Existing fraud detection systems primarily adopt rules based on expert judgment. These systems are unable to identify newly emerging fraud techniques. These inefficiencies are made worse by domain experts' requirement to maintain an extensive collection of intricate rules and labour-intensive manual investigation.

The previous chapter also outlined different supervised machine learning techniques for detecting e-commerce fraud. These techniques use predictive analytics approaches to detect frauds, giving a surprisingly good result. However, a substantial amount of labelled historical data is needed for these supervised learning techniques, which makes it very good at detecting previous fraud patterns but limited in identifying newly emerging fraud techniques.

In addition, different unsupervised machine learning techniques have been explored in detecting fraudulent activities. These techniques are good at detecting various data irregularities, including newly devised fraud techniques. However, these techniques are limited in distinguishing between fraudulent and regular events.

Resolving the deficiency posed by the different techniques requires a novel approach that combines the strengths of each method to improve the fraud detection rate. This research aims at developing a unique e-commerce fraud detection system that can quickly and effectively identify anomalous activities in e-commerce platforms while also dynamically evolving to preserve efficiency with little involvement from subject matter experts. The chapter highlights the description of variables used in this research, the approach used

to perform the study, the research layout, data collection practices and statistical techniques etc.

3.2 Research Approach

The quantitative, qualitative, and mixed methodology is some research methodologies used in a study. Quantitative research focuses on objectivity and is particularly appropriate when it is possible to collect quantitative measures of variables and inferences from population samples by adopting structured procedures and standard tools for data collection (Queirós et al., 2017). Quantitative methodologies include descriptive, experimental, correlational, quasi-experimental and non-experimental research (Rutberg and Bouikidis, 2018). Qualitative research, on the other hand, is not about showing numbers but about deepening the understanding of a particular problem by generating and presenting comprehensive data to understand the different dimensions of the problem under investigation (Queirós et al., 2017). Examples include grounded theory, ethnographic research, historical analysis, case studies, and phenomenology research (Kothari and Garg, 2014). Unlike quantitative research, which helps test hypotheses, predictions, and theorems, qualitative research is helpful for understanding concepts, viewpoints, and opinions (Myers, 2019). This study has quantifiable objectives and emphasises machine learning statistical techniques for data analysis: therefore, it decisively used a quantitative approach.

3.3 Research Framework

This research will combine several strategies, including analytical and empirical techniques, to efficiently address the study objectives. When it becomes important to properly assess and compare alternative ways during the research phase, the analytical technique will be used. The empirical method will be utilized for modelling, cross comparison of the different use-cases, and performance evaluation of the models. Below is the breakdown of this study's structure.

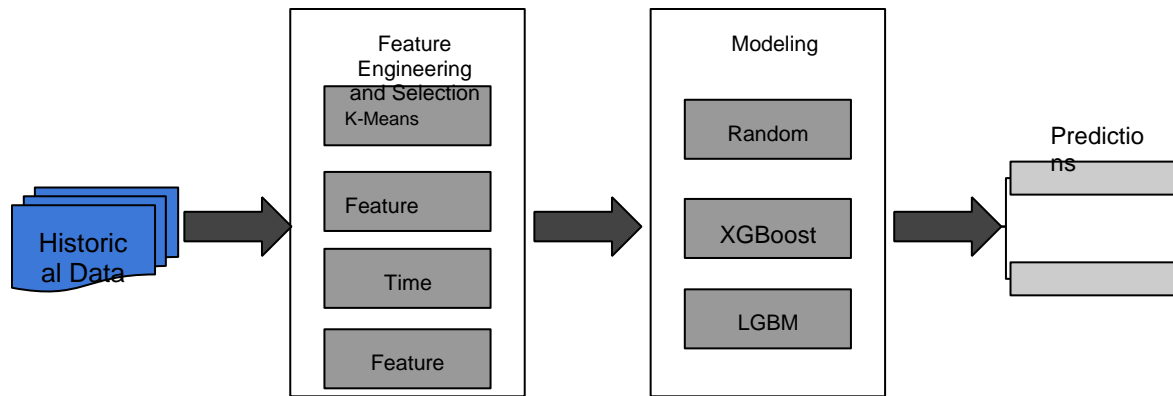


Figure 3.1. Research Framework

The framework for the suggested fraud detection system is shown in the diagram above.

The framework will be composed of the following components:

- Data acquisition: Secondary data will be gathered from a data repository as part of the framework's initial stage, data acquisition, to create a comprehensive customer profile with information that can help the model easily differentiate between fraudulent and non-fraudulent transactions.
- Data preparation and exploration: The collected data will be examined to gather some insights into the fraud patterns and condensed to a more usable form. This stage happens to be one of the essential aspects of successfully training a fraud detection model.
- Feature Engineering: At this stage, the dataset is explored using different techniques to gain more insight into how each feature of the dataset can contribute to the model and how to create more useful features. This will provide deeper insight into fraudulent transactions.
- Model Training and optimisation: At this stage, various machine learning models that can offer a fundamental framework for achieving the aforementioned goals are constructed and contrasted in order to choose the model that performs the best.
- Analysing and Reporting - At this stage, the performance of each model will be evaluated based on some set of selected metrics for the sole purpose of accessing the strength and weaknesses of each model.

The above framework will further be split into two different sections, model development and model evaluation, as shown in Figure 3.2 below.

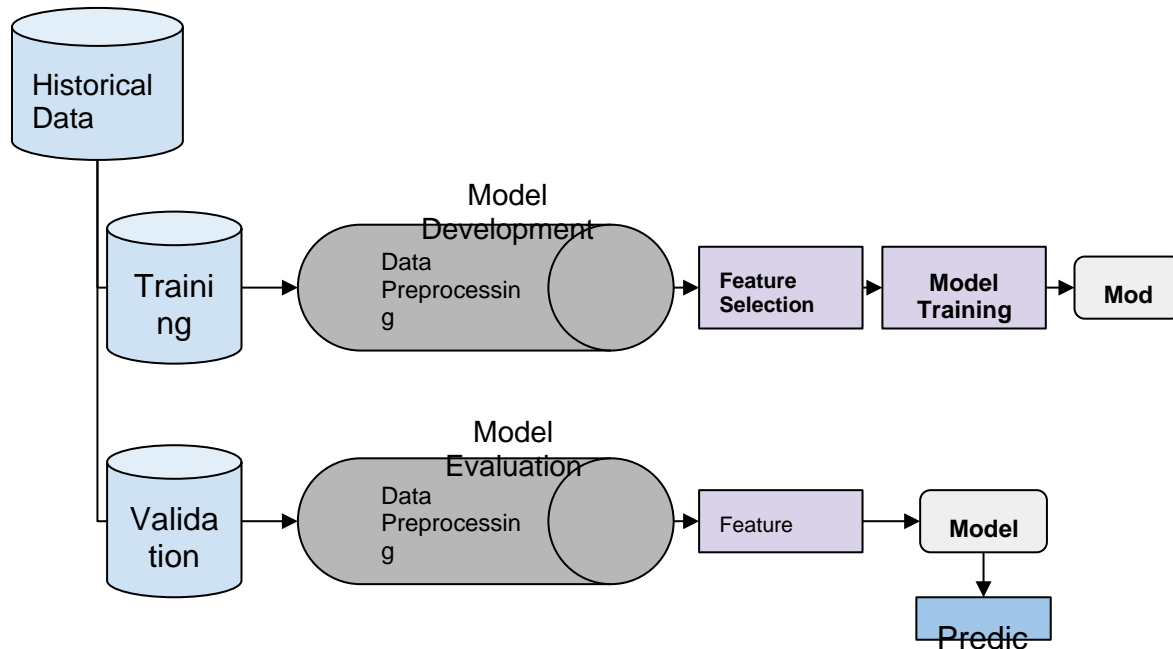


Figure 3.2. Research Framework Sections

The model development section will be responsible for developing the model using historical data. Once the model is trained, it will be passed to the model evaluation section where the models are evaluated. The framework sections will be further expantiated in the subsequent sections.

3.4. Data Collection

3.4.1. Historical Data Collection

In carrying out this research, a secondary dataset will be employed from a reputable source. This study will use an e-commerce fraud dataset sourced from the Kaggle repository. The data come from Vesta's real-world e-commerce transactions database and contains a wide range of features from device type to product features. ("GitHub - chenjing999/IEEE-CIS-Fraud-Detection: fraud detection") Vesta Corporation is the

forerunner in guaranteed e-commerce payment solutions. Founded in 1995, Vesta pioneered the process of fully guaranteed card-not-present (CNP) payment transactions for the telecommunications industry. (“GitHub - Kasyfil97/Fraud-Transaction-Detection-by-Balancing ...”) Two collected data are provided in two different categories:

- I. Transactional information: This dataset contains the information related to the customers' transactions like purchaser and recipient email domain, transaction payment amount, product code, the product for each transaction and others. The dataset contains a large volume of retail transactional events which are labelled as fraudulent or non-fraudulent transactions. The dataset contains 590,540 transactions from different users, with 569,877 normal transactions and 20,663 fraudulent transactions.
- II. Identity information: This dataset contains the customers' identity information – network connection information (IP, ISP, Proxy, etc.) and digital signature (UA/browser/os/version, etc.) associated with their transactions. This dataset contains 144,233 customer information with 41 different features.

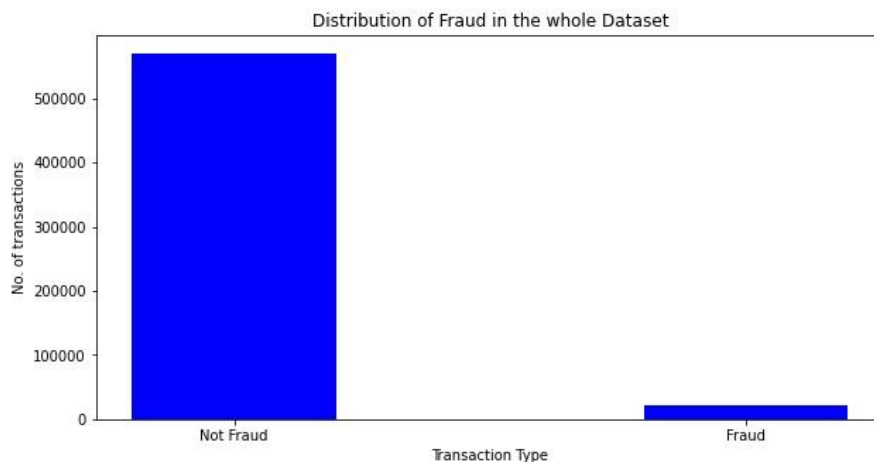


Figure 3.3. Barchart showing the fraud and non-fraud counts

The transactional dataset was labelled by classifying reported chargebacks on the card as fraudulent transactions (isFraud=1) and classifying subsequent transactions that had a direct connection between the user account, email address, or billing address to one of these attributes as fraudulent as well. If none of the aforementioned items is reported or discovered after 120 days, the transaction is deemed legitimate (isFraud=0). Some

fraudulent conduct, however, might not be reported because the cardholder was ignorant or failed to report it in time and after the claim period had passed, for example. In these situations, alleged fraud is presented as legitimate.

Although machine learning methods are a powerful tool for gaining insights from past events, there are some issues that have an impact on the performance of any machine learning classifier in the fraud detection sector. Some of these issues are given as follows:

1. Complex statistical methods are needed to detect fraud using machine learning models, and these methods often demand significant amounts of historical data that aren't always readily available.
2. Predictive modelling can be affected to a large extent by imbalanced dataset classifications because the majority of machine learning methods for classification were built on the premise that there should be an equal number of samples in each class.
3. Due to privacy worries regarding data leaks involving Personally Identifiable Information, businesses are hesitant to publish data in full. As a result, there aren't many credible data sources available.

The above analysis shows that a typical e-commerce fraud detection problem is an imbalance class problem with a high quantity of transactions being genuine transactions and few fraudulent transactions. This happens to be one of the problems affecting the performance of a machine learning model. Therefore, there will be a need to devise a strategy of rebalancing the dataset and ensuring a balanced class of data is maintained during the modelling process. To address the aforementioned issues, a synthetic data creation approach will be implemented.

3.4.2. Synthetic Data Generation

To effectively generate data that are a close representation of the actual data by keeping the statistical relationship intact while increasing the volume of data, the synthetic minority over-sampling technique (SMOTE) will be used in this study. SMOTE maintains the statistical relationship while increasing the volume of data. The objective is to develop

control measures when it is challenging for a machine learning system to distinguish between accurate and generated data given a set of conditions to create new data.

This method looks at several specific events that users may carry out on a retail channel. After that, synthetic data for the low occurrence class, fraudulent transactions, is produced utilizing this data, which is then used for training and evaluation. This method seeks to alleviate the overfitting issue so that models can generalize more effectively by maintaining the statistical inference qualities of the original data and producing extra data that has a nature comparable to the original within a specific bound of error confidence.

3.5 Feature Engineering

The core of this research will be feature engineering. This procedure will produce features that the suggested models can process using domain expertise. Here, the data will be examined to discover how different data features affect the model performance.

The feature engineering process will start by gaining insight into the characteristics of the data needed for the different machine models by relating them to the challenge of fraud detection. Here, the effects of variables like device data, purchase quantity, transaction amount, behaviour data, and others will be studied.

More features are then created and validated by running experiments to better understand the dataset and how different customer interaction features can help improve the model performance.

3.6. Model Description

In this study, three ensemble machine learning models for fraud detection will be developed, i.e., Random Forest Classifier as a Bagging Model, Extreme Gradient Boosting

(XGBoost) and Light Gradient Boosting (LGBM) Models as a type of Boosting Model. In addition to these three models, features will be generated from an Unsupervised machine learning model, the K-Means Clustering algorithm, and fed to the ensembling machine learning models to improve their performance. The various categories of the models to be implemented will be described below:

3.6.1 Clustering Algorithm

Clustering algorithms are one of the significant subsets of unsupervised learning. The clustering method seeks to identify clusters of related observations within the data that are characterized by the highest degree of similarity within a group and the highest degree of dissimilarity across groups (Sinaga and Yang, 2020). Clustering can be used in retail to identify both regular and fraudulent transactions. Customers' sociodemographic, behavioural, or other account characteristics, like a customer transaction, can be used to derive the different clusters in the dataset. This study will use the K-Means clustering model as a feature engineering technique to create features that can help the other supervised machine learning models to capture different customer segmentation present in the dataset and in detecting various data irregularities, including newly devised fraud techniques.

3.6.2 Bagging Models.

Bagging predictors is a technique for creating numerous types of a predictor and using these to produce an aggregated predictor. When forecasting a numerical conclusion, the aggregate takes an average across all variants, and when predicting a class, it performs a plurality vote. The different versions are created by creating bootstrap copies of the learning set and using these as brand-new learning sets. Bagging can result in significant improvements in accuracy, according to tests on actual and simulated data sets utilizing classification and regression trees, as well as subset selection in linear regression. The prediction method's instability is a crucial component. Bagging can increase accuracy if perturbing the learning set can result in noticeable changes to the predictor that was built.

A machine learning ensemble meta-algorithm called bagging is intended to improve the accuracy and stability of machine learning algorithms. The technique minimizes variance, which impacts the model's performance and helps prevent overfitting. The Bagging Model implemented in this chapter is the Random Forest Classifier.

3.6.3 Boosting Models

Boosting algorithms are among the most promising methodological developments in the last 20 years for data analysis (Mayr et al., 2014). The fundamental concept is to apply weak classifiers iteratively and aggregate the results to produce more accurate predictions. Compared to previous models, the LightGBM model, according to Eom et al., (2021), achieves a significantly better result. The approach also has a better reach for evaluation and optimization due to the cross-validation benefit of ensemble models on each weak classifying model. Two types of boosting models, XGBoost and LighGBM will be implemented in this research.

3.7. Performance Evaluation Metrics

This section outlines the various methods for determining how effective a machine learning model is. Measuring the expectation ratio using methods is a standard way to evaluate the effectiveness of the framework. This ratio includes:

- I. True Positive Rate (TPR): This is the rate of non-fraudulent transactions correctly classified as non-fraudulent.
- II. False Positive Rate (FPR): This is the rate of fraudulent transactions misclassified as non-fraudulent transactions.
- III. True Negative Rate (TNR): This is the rate of non-fraudulent transactions misclassified as fraudulent transactions.
- IV. False Negative Rate (FNR): This is the rate of fraudulent transactions correctly classified as fraudulent transactions.

The confusion matrix as shown in Table 3.18 below gives an overview of the metrics:

		True class (y)	
		Fraud (y = 1)	Not Fraud (y = 0)
Predicted Class (\hat{y})	Fraud ($\hat{y} = 1$)	TP (True Positive)	FP (False Positive)
	Not Fraud ($\hat{y} = 0$)	FN (False Negative)	TN (True Negative)

Table 3.1. Confusion Matrix

The metrics shown in the table above on their own, aren't enough for evaluation. They are a baseline for some other useful metrics. The following are some of the metrics that can be calculated using the above metrics as baseline:

$$\text{Accuracy score} = \frac{TP + TN}{TP + TN + FP + FN} \quad (\text{eq 3.1})$$

$$\text{Error rate} = 1 - \frac{TP + TN}{TP + TN + FP + FN} \quad (\text{eq 3.2})$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (\text{eq 3.2})$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (\text{eq 3.4})$$

$$\text{F1 score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (\text{eq 3.5})$$

$$\text{True Positive Rate} = \frac{TP}{TP + FN} \quad (\text{eq 3.6})$$

$$\text{False Positive Rate} = \frac{FP}{TN + FP} \quad (\text{eq 3.7})$$

The aim of the models is to reduce the False Negative Rate and False Positive Rate as much as possible and to increase the True Negative Rate and True Positive Rate as much as possible. (“Algorithms | Free Full-Text | Extraction and Segmentation of ... - MDPI”) To better assess the performance of the models, F1 score will be the preferred evaluation metric.

3.8. Baseline Model

To evaluate the effectiveness of the models to a well implemented supervised machine learning technique in ecommerce fraud detection, a baseline model will be needed. A Support Vector Machine (SVM) Classifier will be trained and evaluated in this section as it happens to be one of the popular machine learning techniques implemented in fraud detection. The SVM model generates a decision boundary in an increased or infinitedimensional space, which is suitable for non-linear classification problems (Naveen, P. and Diwan, B., 2020). The confusion matrix below shows the performance of the baseline model.

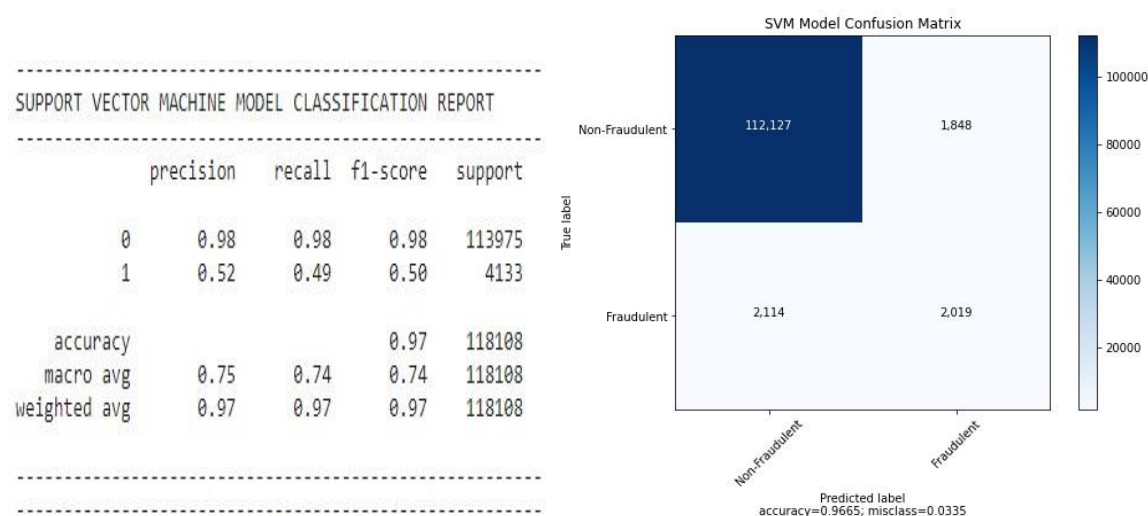


Figure 3.4 SVM Model Evaluation Report

As shown in Figure 3.5 above, the SVM model correctly labels most non-fraudulent transactions (99%) but incorrectly labels most fraud cases (51%). This indicates that the

model has a false positive rate of 90%. The goal of training this model is to compare the performance of SVM with other models based on the same dataset.

3.9. Conclusion

The primary components of this study—namely, feature engineering, assembling machine learning models, and model evaluation—are covered in this chapter. This chapter also described a framework for fraud detection. The chapter's proposed feature engineering approach which focuses on taking transactional, customer behavioural, and personal data and creating new features the ensemble classifiers can use. The implementation methodology is presented in the next chapter, enabling us to implement several experiments on the data using the Random Forest Classifier, Extreme Gradient Boosting, and other ensembling models. The results of each experiment will be computed using the selected performance metric to determine how the five algorithms fare when used to solve a challenging real-world situation.

3.10. Summary of Procedure

The summary of this chapter's material is described in this section.:

1. It proposed a novel fraud detection system that combines supervised machine learning techniques with ensemble classifiers to identify fraudulent transactions effectively.
2. Details of the dataset selected for modelling.
3. Data feature engineering and selection techniques emphasise developing new features that can prove helpful to the models.
4. Provides insight into the different ensembling machine learning models to be implemented in this study
5. Presents different evaluation metrics for model performance.

CHAPTER FOUR DATA ANALYSIS AND MODEL IMPLEMENTATION

4.1. Model development

This chapter provides a detailed implementation of the selected models on a real-life ecommerce dataset from Vesta Corporation. Based on feature engineering and technique that combines machine learning models, the model can constantly evolve to retain efficiency with little input. The model consists of four primary modules: Data Preprocessing, Feature Engineering, Model Implementation, Ensemble Models and Model Evaluation. The Feature Engineering section examines and analyzes the transactional events gathered from an e-commerce database. Features are then created to quickly construct a feature matrix that fraud detection models can handle soon. In addition, training and testing sample datasets are gotten from the original dataset. The models classify new transactions by learning the patterns in customers' historical transactional datasets. The process of training the model in this study involves two main steps; first, features are created using different feature engineering techniques, with a particular one coming from the clustering of different customer transactional information to create different clusters that can provide insights and patterns which are not readily available in the dataset. These features are then fed to the ensemble models to produce a better classification algorithm.

4.2. Data Preprocessing

4.2.1. Data Description

The data is obtained from Vesta's real-world e-commerce transactions database and contains a wide range of features from device type to product features. ("Smart Surgical Assistance - AI Cases") This dataset includes data about the customer's transactions and identity information like purchaser and recipient email domain, transaction payment amount, product code, the product for each transaction, and network connection information (IP, ISP, Proxy, etc.) and others. The dataset contains many retailed



transactional events labelled as fraudulent or nonfraudulent transactions. The dataset includes 590,540 transactions from different users, with 569,877 everyday transactions and 20,663 fraudulent transactions. The dataset is highly imbalanced, with fraudulent transactions accounting for only 3.5% of the total transactions. A quick snapshot of the data is shown in Figure 4.1 below.

Transactions Data shape : (590540, 394)
Memory usage of dataframe is 1775.15 MB
Memory usage after optimization is: 487.16 MB
Decreased by 72.6%

1 to 5 of 5 entries  

index	TransactionID	isFraud	TransactionDT	TransactionAmt	ProductCD	card1	card2	card3	card4	card5	card6	addr1	addr2	dist1
0	2987000	0	86400	68.5	W	13926	NaN	150.0	discover	142.0	credit	315.0	87.0	19.0
1	2987001	0	86401	29.0	W	2755	404.0	150.0	mastercard	102.0	credit	325.0	87.0	NaN
2	2987002	0	86469	59.0	W	4663	490.0	150.0	visa	166.0	debit	330.0	87.0	287.0
3	2987003	0	86499	50.0	W	18132	567.0	150.0	mastercard	117.0	debit	476.0	87.0	NaN
4	2987004	0	86506	50.0	H	4497	514.0	150.0	mastercard	102.0	credit	420.0	87.0	NaN

Identity Data shape : (144233, 41)
Memory usage of dataframe is 45.12 MB
Memory usage after optimization is: 10.00 MB
Decreased by 77.8%

1 to 5 of 5 entries  

index	TransactionID	id_01	id_02	id_03	id_04	id_05	id_06	id_07	id_08	id_09	id_10	id_11	id_12	id_13	id_14	id_15	id_16
0	2987004	0.0	70787.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	100.0	NotFound	NaN	-480.0	New	NotFound
1	2987008	-5.0	98945.0	NaN	NaN	0.0	-5.0	NaN	NaN	NaN	NaN	100.0	NotFound	49.0	-300.0	New	NotFound
2	2987010	-5.0	191631.0	0.0	0.0	0.0	0.0	NaN	NaN	0.0	0.0	100.0	NotFound	52.0	NaN	Found	Found
3	2987011	-5.0	221832.0	NaN	NaN	0.0	-6.0	NaN	NaN	NaN	NaN	100.0	NotFound	52.0	NaN	New	NotFound
4	2987016	0.0	7460.0	0.0	0.0	1.0	0.0	NaN	NaN	0.0	0.0	100.0	NotFound	NaN	-300.0	Found	Found

Figure 4.1. Dataset snapshot

Transaction dataset features:

Features	Description	Data Type
TransactionDT	Timedelta from a given reference datetime (not an actual timestamp)	Categorical
TransactionAMT	Transaction payment amount in USD	Numerical

ProductCD	Product code, the product for each transaction	Categorical
card1 - card6	Payment card information, such as card type, card category, issue bank, country, etc.	Categorical
addr	Address	Categorical
dist	Distance	
P_emaildomain and R_emaildomain	Purchaser and recipient email domain	Categorical
C1-C14	Counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.	Categorical
D1-D15	Timedelta, such as days between previous transactions, etc.	Categorical
M1-M9	Match, such as names on card and address, etc.	Categorical
V1-V399	Vesta engineered rich features, including ranking, counting, and other entity relations	Numerical

Table 4.1. Transaction Dataset Features

Identity dataset features:

Variables in the identity dataset are identity information – network connection information (IP, ISP, Proxy, etc.) and digital signature (UA/browser/os/version, etc.) associated with transactions. The field names are masked for privacy protection and contract agreement.

These features include:

- DeviceType (Categorical): The type of device used for the transaction.
- DeviceInfo (Categorical): Information about the device used.
- Id_12 - Id_38: This contains device rating, ip_domain rating, proxy rating, account login times/failed to login times, how long an account stayed on the page, etc.

The datasets provide a series of customer events from the point at which they start the transaction to the point at which the transactions are completed. The transaction dataset contains 394 features, while the identity dataset contains 41 features, totalling 434 features after merging them on the TransactionID feature. Some of the features with missing values are either eliminated or filled with descriptive values, such as the most recent value for categorical features and the average for numerical features, during the modelling process to reduce the dimensionality and bias of the data. After the data preprocessing, a total of 131 features were selected from the dataset. The data selection and preprocessing are further illustrated in more detail in the preprocessing data subsection.

4.2.2. Exploratory Data Analysis

This section's exploratory analysis aims to comprehend and establish the distributions of the datasets and the primary data features on which the model relies. The dataset was splitted into training and validation by selecting 20% of the overall dataset as validation dataset and the remaining 80% as training dataset. The dataset was split using a stratified sampling technique to keep the ratio of fraudulent to non-fraudulent transactions the same across the splits. The training dataset contains 455,902 non-fraudulent transactions

and 16,530 fraudulent transactions, while the validation dataset, on the other hand, contains 113,975 non-fraudulent transactions and 4,133 fraudulent transactions.

The numerical analysis of the training and validation datasets above shows that the fraud ratio for both datasets is approximately the same, 3.5%, the same as the overall dataset.

	Fraudulent	Non-Fraudulent
Training Dataset	16,530	455,902
Validation Dataset	4,133	113,975
Overall Dataset	20,663	569,877

Table 4.2. Distributions of the fraudulent and non-fraudulent transactions



Figure 4.2. Distributions of fraudulent and non-fraudulent transactions

Transaction Amount and the transaction date and time features are employed in this analysis. The validation dataset is ignored, and only the training dataset is examined.

The distribution of non-fraudulent and fraudulent transactions, as shown in Figures 4.3 and 4.4, shows that transactions at lower amounts, less than \$300, are common for both non-fraudulent and fraudulent transactions. The distributions of non-fraudulent transactions are more even than those of fraud transactions, demonstrating that fraud occurs less frequently. As shown in the two distributions, the range of transaction

amounts is between \$1 and \$900. This is not the actual range from the dataset. It is only trimmed to that particular range since the majority of transactions happen to fall below \$900, with a small transaction amount greater than \$900. The distribution plot in Figure 4.5 is generated to better show the whole range.

Figure 4.5 shows the distribution of fraudulent and non-fraudulent by transaction amounts range. Most fraudulent transactions had payments ranging from \$1 to \$600. However, some fraudulent transactions with big sums up to \$1,000 occurred. In addition to Figure 4.5, the frequency of both fraudulent and non-fraudulent transactions and the ranges of transaction amounts are shown in Table 4.3 below.

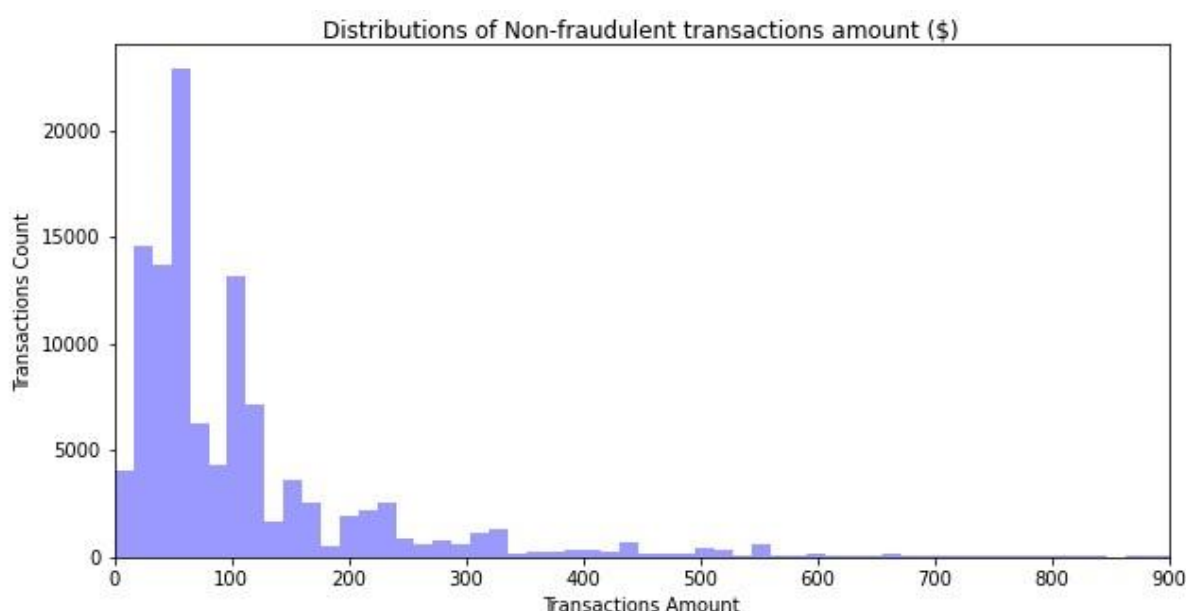


Figure 4.3. Distributions of Non-fraudulent transactions

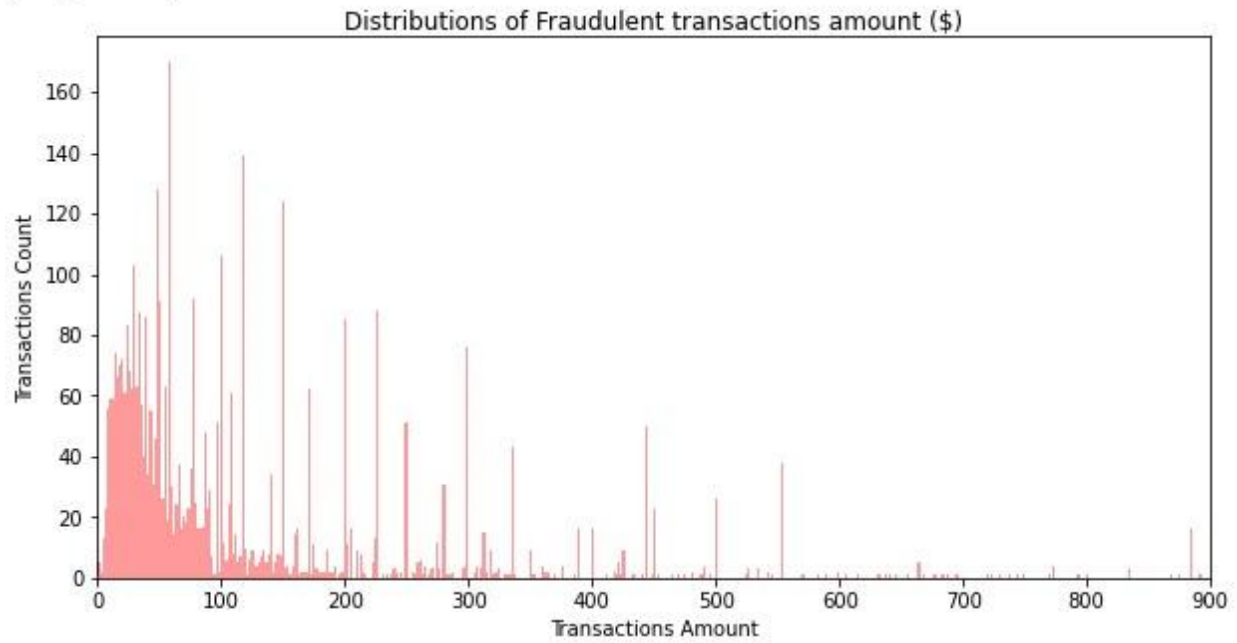


Figure 4.4. Distributions of Fraudulent transactions

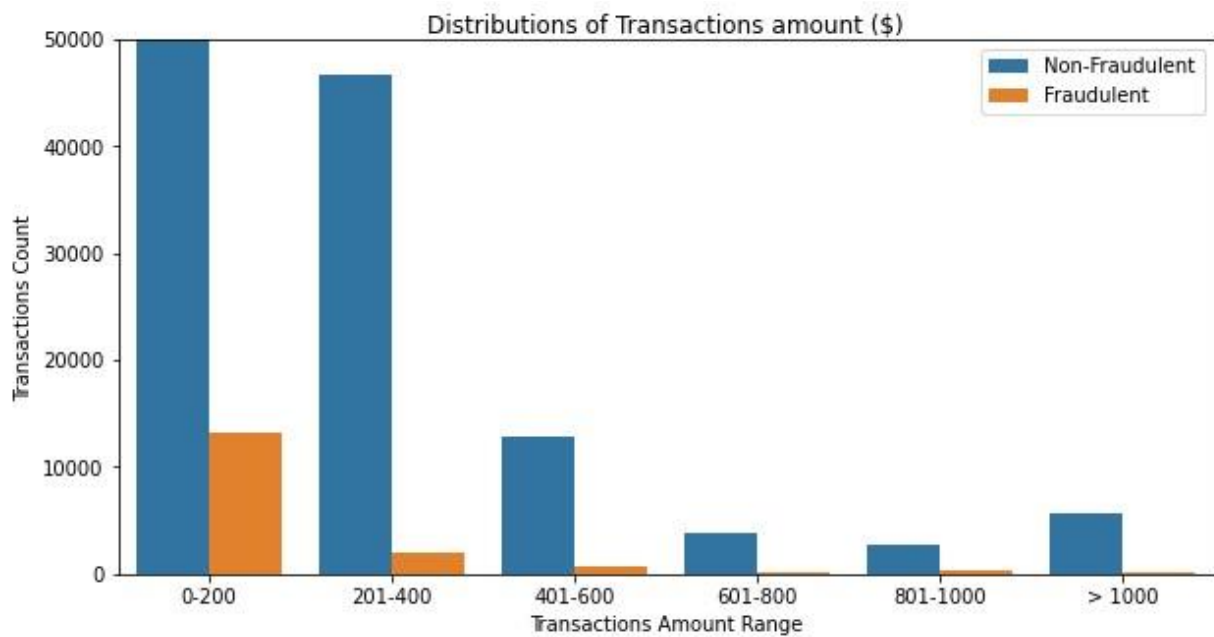


Figure 4.5. Number of Fraudulent and Non-Fraudulent Transactions by Amount Range

Transaction Amount	Non-Fraudulent	Fraudulent
0-200	384,262	13,234
201-400	46,616	1,945
401-600	12,821	740
601-800	3,864	159
801-1000	2,656	315
> 1000	5,629	135

Table 4.3. Number of Fraudulent and Non-Fraudulent Transactions by Amount Range

Figure 4.6 and Figure 4.7 overlay the number of fraudulent and non-fraudulent transactions by time, with Figure 4.6 measured per hour and Figure 4.7 measured in days of the week. In the figures, the orange bars represent fraudulent transactions, and the blue bars represent non-fraudulent transactions. Figure 4.6 shows that both transactions occur more often during the day, with the fraud rate being approximately equal across the hour. Figure 4.7 shows that the fraud rate is roughly equal across the day of the week.

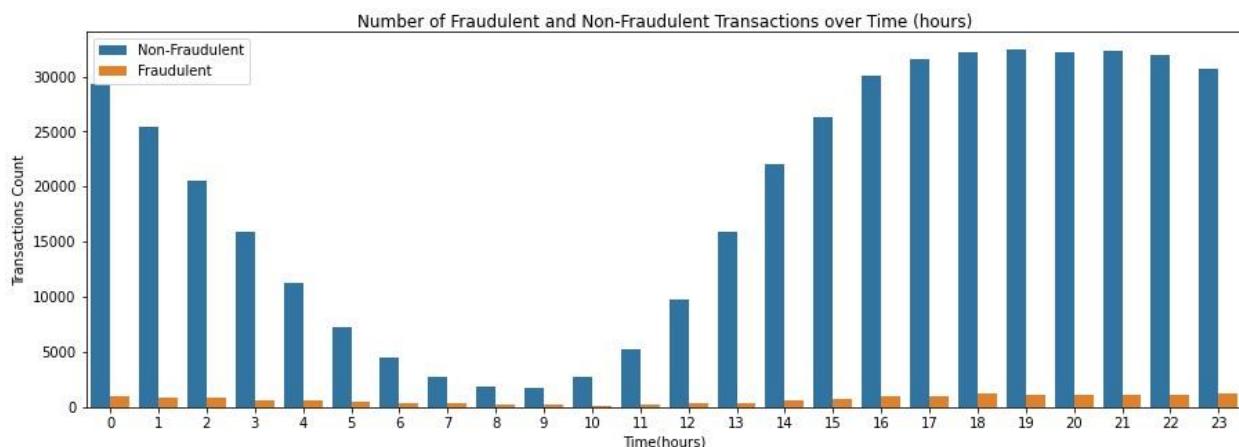


Figure 4.6. Number of Fraudulent and Non-Fraudulent Transactions by hour

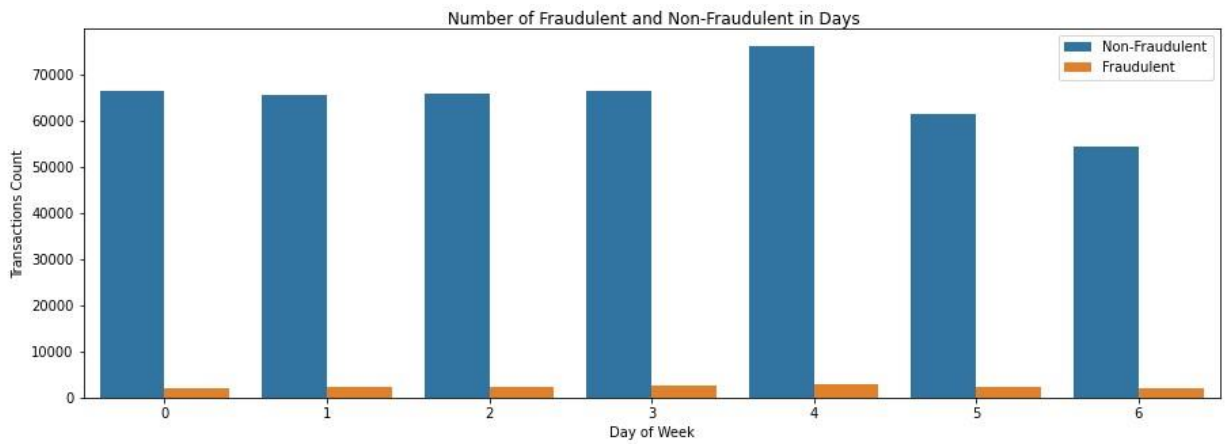


Figure 4.7. Number of Fraudulent and Non-Fraudulent Transactions by Day of week

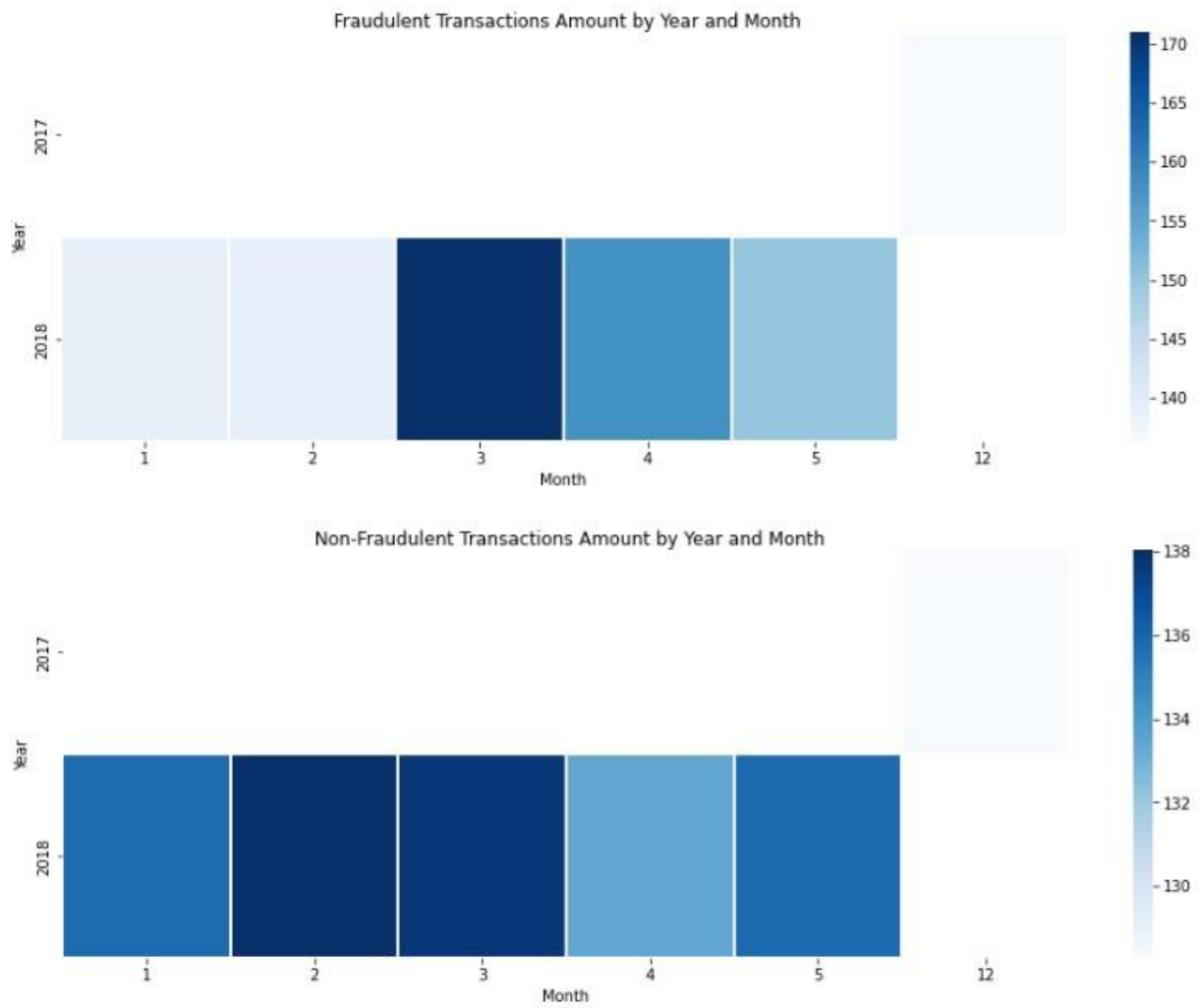


Figure 4.8. Transactions Amount by Year and Month

Figure 4.8 shows the heat map of the non-fraudulent and fraudulent transactions amount grouped by year and month. The first figure shows that the highest transaction amount for fraudulent transactions was made in March 2018, while the second figure shows that the highest transaction amount for non-fraudulent transactions was made in March and February 2018.

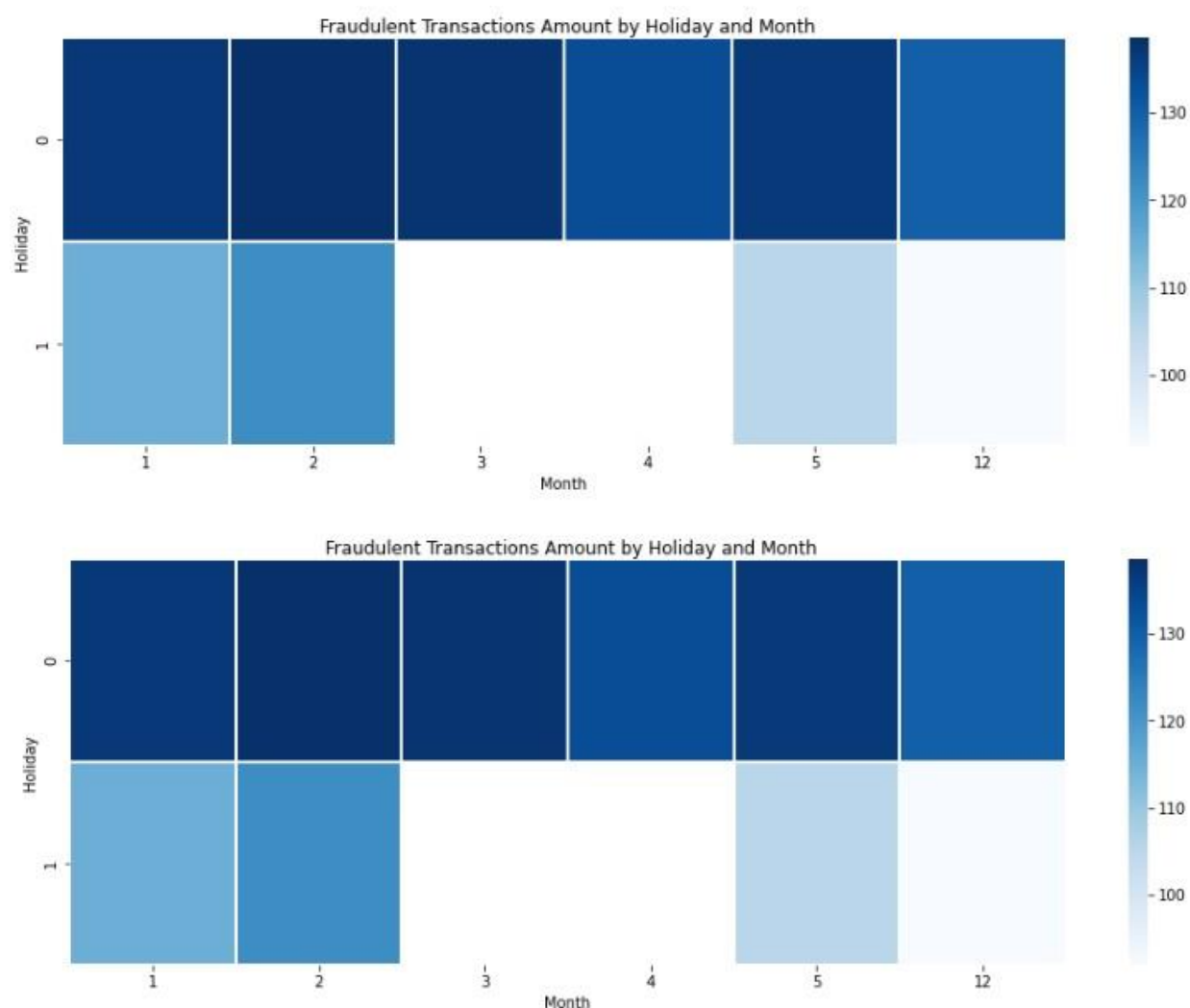


Figure 4.9. Transactions Amount by Holiday and Month

Figure 4.9 shows the heat map of the non-fraudulent and fraudulent transactions amount grouped by Holiday and month. The figures shows approximately the same distribution for the fraudulent and Non-fraudulent transactions, with majority of the transaction payment occurring at the months without

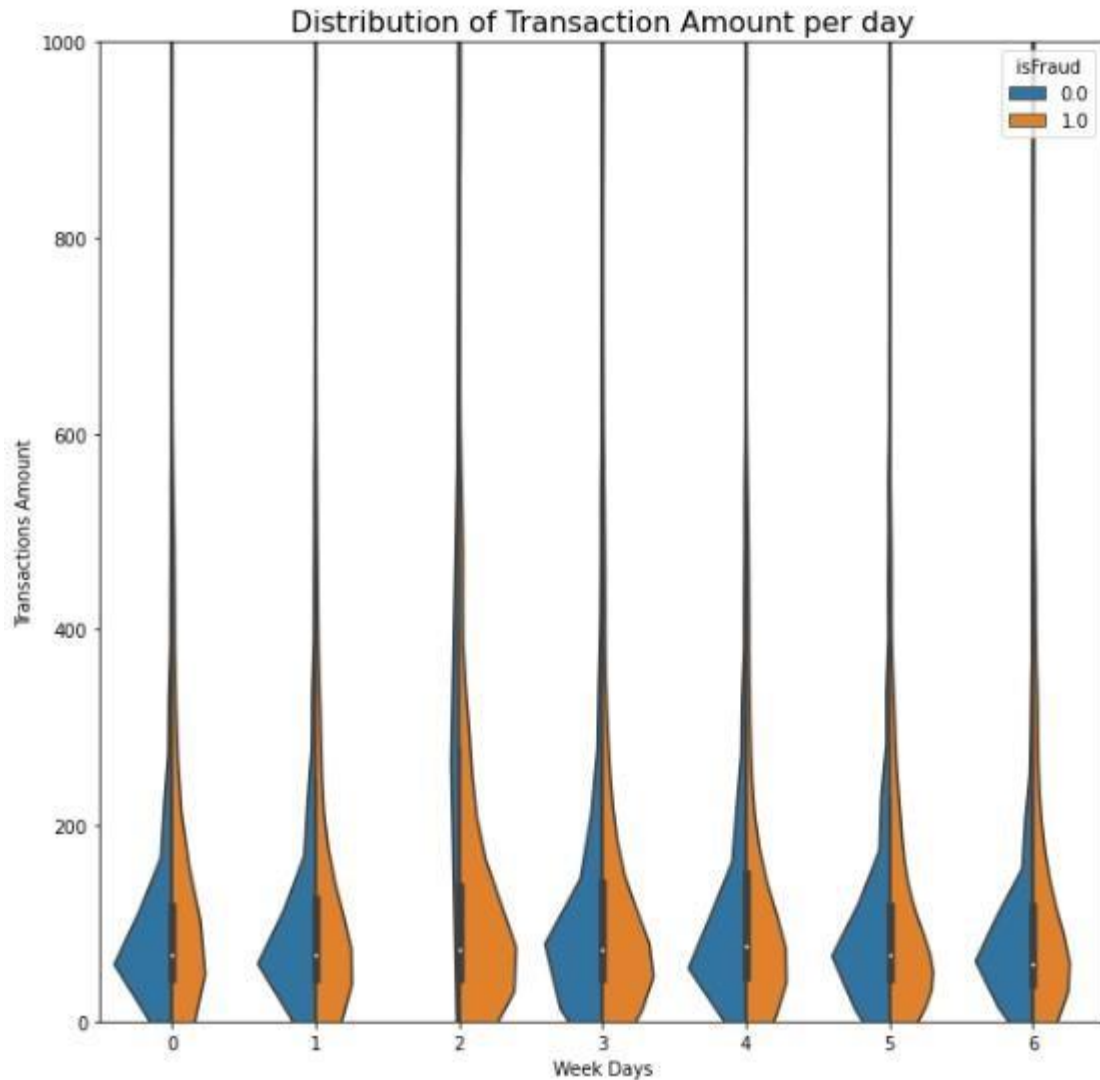


Figure 4.10. Transactions Amount Distribution by Weekdays

Figure 4.10. shows the distribution of fraudulent and non-fraudulent transaction amounts by weekdays. In the figures, the orange bars represent fraudulent transactions, and the blue bars represent non-fraudulent transactions. This plot shows that the distribution of the fraudulent and not fraudulent transactions across each weekdays are approximately constant.

4.3 Feature Engineering

Feature engineering is a crucial step in preparing data, and it helps improve the performance of machine learning models by creating appropriate features from provided features. In feature engineering, existing components are subjected to transformation operations like arithmetic and aggregation operators to develop new ones. With the use of transformations, a feature can be scaled, or a non-linear relationship between a component and a target class can be changed into a linear relation, which is simpler to understand. Different feature engineering was performed by following the feature engineering framework provided in chapter 3. This process is an iterative creation validation process where features are created, tested and selected or discarded based on how well the model performs with the feature. The different feature engineering techniques used in this study are given below:

4.3.1 Feature Engineering Approaches

1) Feature Binning

Feature binning will be performed on the dataset because it contains so many features, of which some are non-linear. The predictive accuracy of any model, in particular ensemble models like the random forest, lightgbm and xgboost, is limited by the noise generated by these features. Binning these features reduces observational error and makes it easier to spot outliers in the dataset. Therefore, the categorical data can be simply normalized into numerical data using selected encoding techniques.

2) Feature interactions

Feature interaction involves the integration of two or more features to create better features. Feature interaction was performed on the numerical features by taking the ratio of two features while the categorical features were concatenated and then encoded to form new features.

3) K-Mean Clustering

K-Means clustering algorithm is applied to the dataset's features to create different clusters. Clustering techniques are good at detecting various data irregularities, including newly devised fraud techniques which can help improve the accuracy of the supervised learning model. The features are split into three mutually exclusive categories, and then the K-Means algorithm is applied to each type to create clusters of different sizes. The cluster size of each class is chosen in such a way as to reduce the error rate of each model. The distribution of each created cluster for the number of fraudulent and non-fraudulent transactions is shown in Figure 4.1 below.

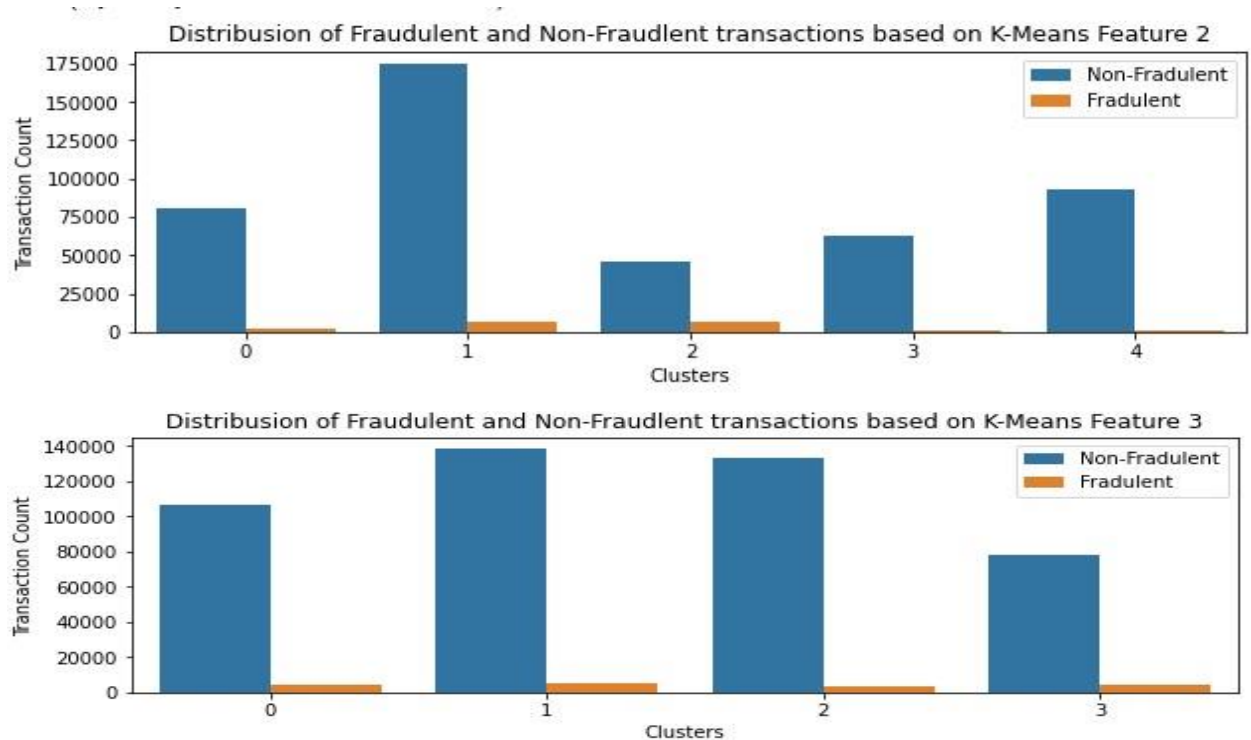


Figure 4.11. Number of Fraudulent and Non-Fraudulent Transactions by clustering

4) Time Series Features

Different features are created from the transaction date time feature. The features developed in this section include the transaction day of the week, transaction

month, transaction hour, transaction day of month and others. These features are necessary as some fraud events sometimes are skewed toward a time.

5) Data resampling

A sampling technique is applied to improve the lower class to overcome the imbalance state of the dataset. Synthetic Minority Over-sampling Technique (SMOTE) was used to increase the number of fraudulent transactions in the dataset. The number of non-fraudulent was then down-sized. This process improved the class imbalance by increasing the ratio of fraudulent transactions in the training dataset from 3.5% to 23.1%.



Figure 4.12. Distributions of the fraudulent and non-fraudulent transactions after sampling

6) Categorical Encoding

Encoding is a crucial factor to consider while modelling and converting categorical variables to numerical ones. Out of all the encoding methods, weights of occurrence were used to represent each category according to how frequently it appears in the dataset.

4.4. Model Training and Implementation

Four ensemble machine learning models for fraud detection will be developed i.e. Random Forest Classifier as a type of Bagging Model, Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LGBM) Models as a type of Boosting Model. The implementation of the models are described below:

4.4.1 Random Forest Model

Random forest is a popular Supervised Machine Learning Algorithm for Classification and Regression issues. It creates decision trees from multiple samples and applies regression using the mean and classification using the majority vote (Xuan, S. 2018). Various trained decision tree models, such as the random forest model, are used to form an ensemble technique. By averaging the output of many models created from randomized data creation, random forest corrects for the overfitting feature of decision trees.

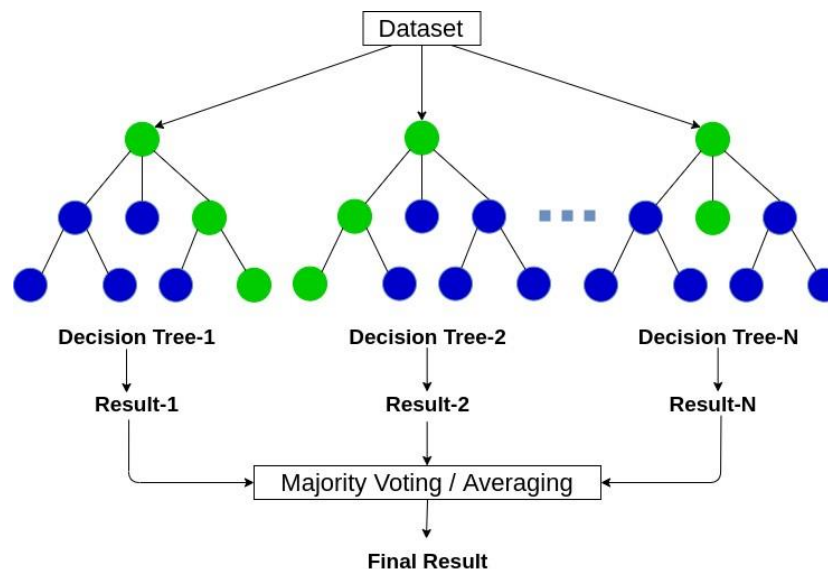


Figure 4.13. Flow of Random Forest Algorithm Operations

As shown in Figure 4.10 above, the Random Forest algorithm randomly selects samples from a given dataset and then constructs a decision tree for every sample. The various decision trees then return a prediction on which voting is performed. The last stage

involves the selection of the most voted prediction as the final prediction. Due to the diversity of the bagging feature and the step-wise aggregation of the different subsets, this model has the advantage of being immune to the curse of dimensionality, which is the main argument for its implementation. The combined forecast considers the bias, which improves the model's stability.

4.4.2 Gradient Boosting Models

Gradient tree boosting is a machine learning method that excels in various practical applications. On many standard classification benchmarks, tree boosting has been demonstrated to produce cutting-edge results by reducing the bias and variance of a dataset. Boosting aids in turning weak learners into strong ones by combining several weak classifiers in series to develop a model. Strong learners correlate strongly with the correct classification, whereas weak learners have a moderate correlation. First, a model is created using the training set data. The second model is then created to fix the previous model's error. Models are added in this manner until either the entire training data set is adequately predicted or the number of models added has reached its maximum.

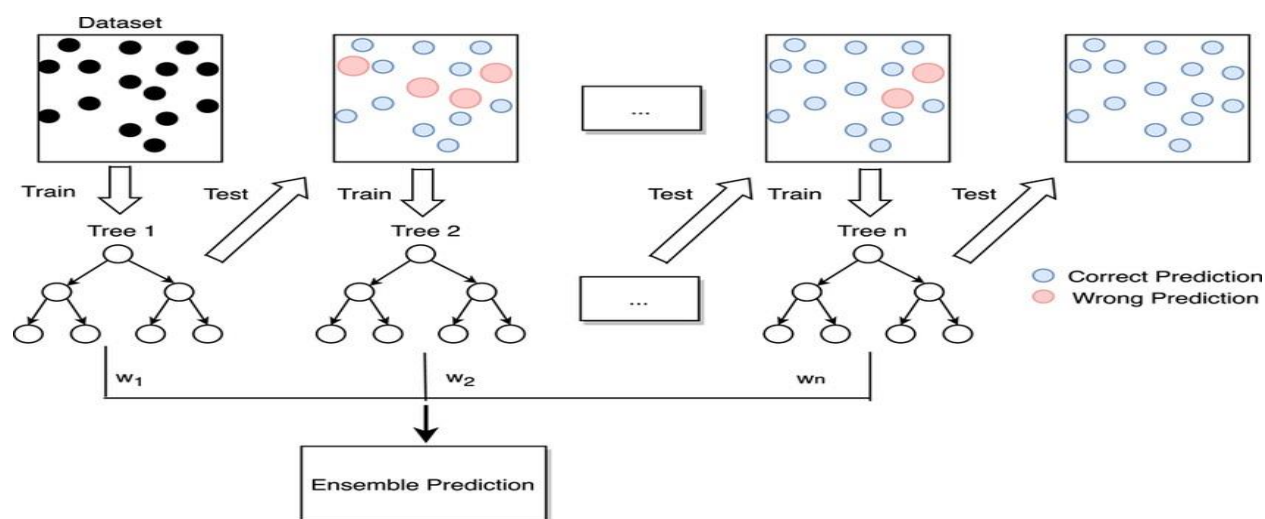


Fig 4.14. Gradient Boosting Algorithm Operations

XGBoost implements Gradient Boosted decision trees, in which decision trees are generated sequentially. Each independent variable is given weight before being fed into the decision tree. Variables that the tree incorrectly predicted are given more weight

before being placed into the second decision tree. These distinct classifiers/predictors are combined to produce a robust and accurate model. XGBoost was primarily created using gradient-boosted decision trees for speed and performance. XGBoost, also known as extreme gradient boosting, aids in using all available hardware and memory resources for tree boosting algorithms.

The LightGBM framework uses gradient-boosting decision trees to improve model performance using less memory. Like XGBoost, the LightGBM improves the gradient boosting method by including an automatic feature selection method and concentrating on boosting instances with more significant gradients. As a result, training can go significantly faster and improve prediction accuracy. Scalability and performance are the primary concerns of this model. It employs two cutting-edge methods: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB), which overcomes the drawbacks of the histogram-based approach. These two methods enable the model to function effectively and gain an advantage over competing GBDT (Gradient Boosting Decision Tree) frameworks.

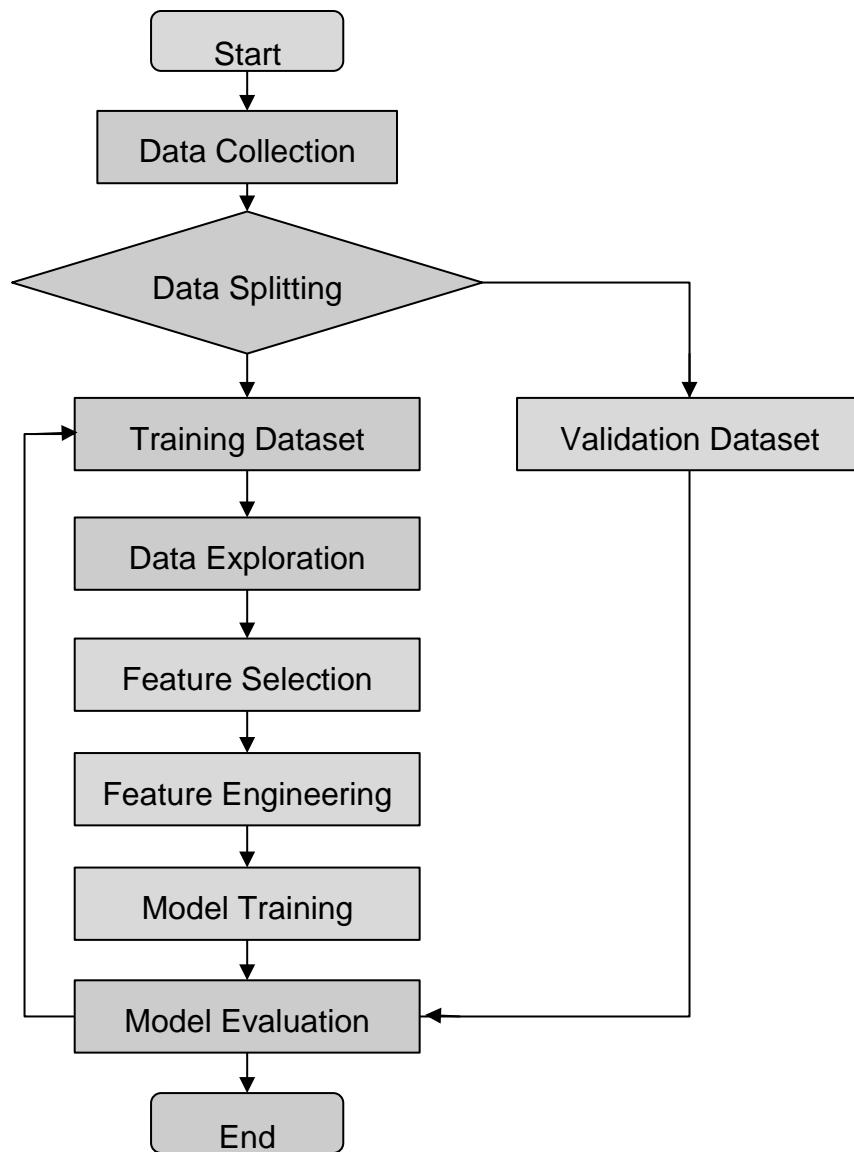


Figure 4.15. Fraud Detection Modeling Process

CHAPTER FIVE EVALUATION AND VALIDATION

5.1. Modelling Results and Analysis

Table 5.1. below shows the number of non-fraudulent and fraudulent transactions in the training and validation dataset after applying the SMOTE technique to balance the data between the two classes.

	Fraudulent	Non-Fraudulent
Training Dataset	45,590	151,966
Validation Dataset	4,133	113,975

Table 5.1. Datasets split into training and validation

Features selection was performed by feeding all the dataset features to each model and then using their in-built feature importance function to filter out the most important features for the model. The computation of the Receiver Operating Characteristic (ROC) curves and F1 score serves as the foundation for evaluating the outcomes of each model.

5.1.1. XGBoost Model Performance

Figure 5.1. below shows the performance of the extreme gradient boosting model on the validation dataset. As shown in the classification report, the model performs better on non-fraudulent events, with an F1 score of 99%. The F1 score on the fraudulent events is 76%. Out of the 4,133 fraudulent transactions, 3,146 were correctly predicted as fraudulent, while the other 987 were incorrectly predicted as non-fraudulent. The overall F1 score of the model is approximately 75.9% which shows the model, on a general note, is pretty good at differentiating the fraudulent and non-fraudulent transactions.

XGBOOST CLASSIFICATION REPORT				
	precision	recall	f1-score	support
0	0.99	0.99	0.99	113975
1	0.76	0.76	0.76	4133
accuracy			0.98	118108
macro avg	0.87	0.88	0.88	118108
weighted avg	0.98	0.98	0.98	118108

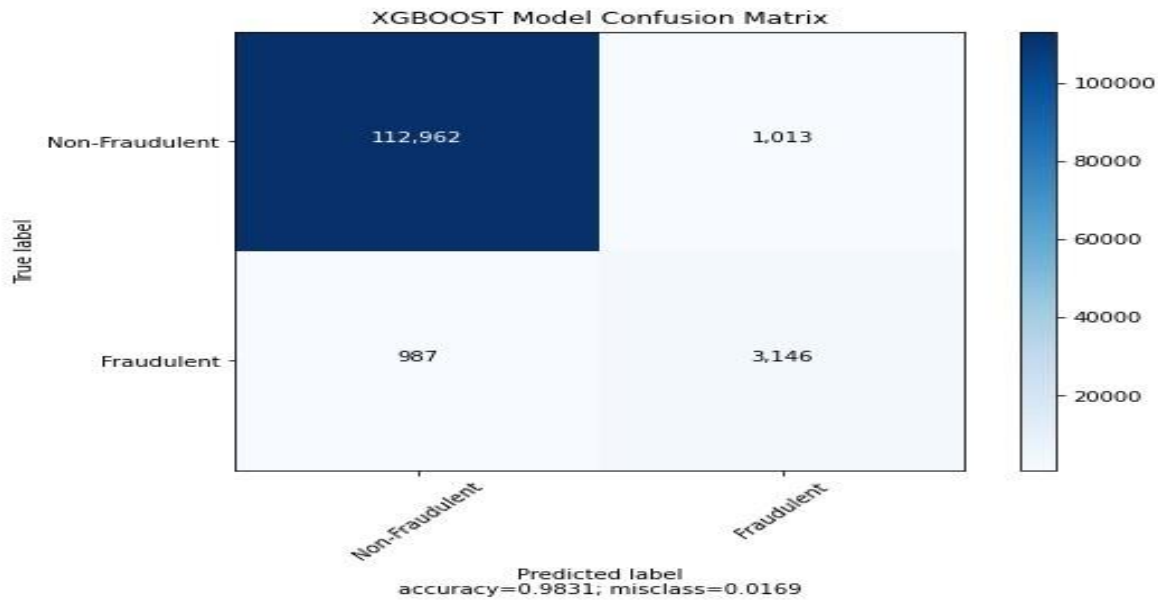
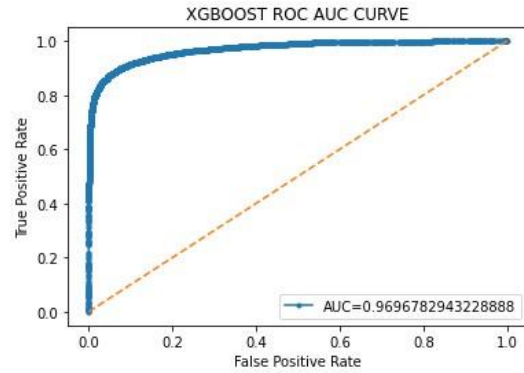


Figure 5.1. XGBoost Model Evaluation Report

5.1.2. LGBM Model Performance

Figure 5.2. below shows the performance of the LGBM model on the validation dataset. As shown in the classification report, the model performs better on non-fraudulent events, with an F1 score of 98%. The F1 score on the fraudulent events is 56%. Out of the 4,133 fraudulent transactions, 3,165 were correctly predicted as fraudulent, while the other 968 were incorrectly predicted as non-fraudulent. The overall F1 score of the model is approximately 55.9% which shows the model, on a general note, is not really good compared to the XGBoost model.

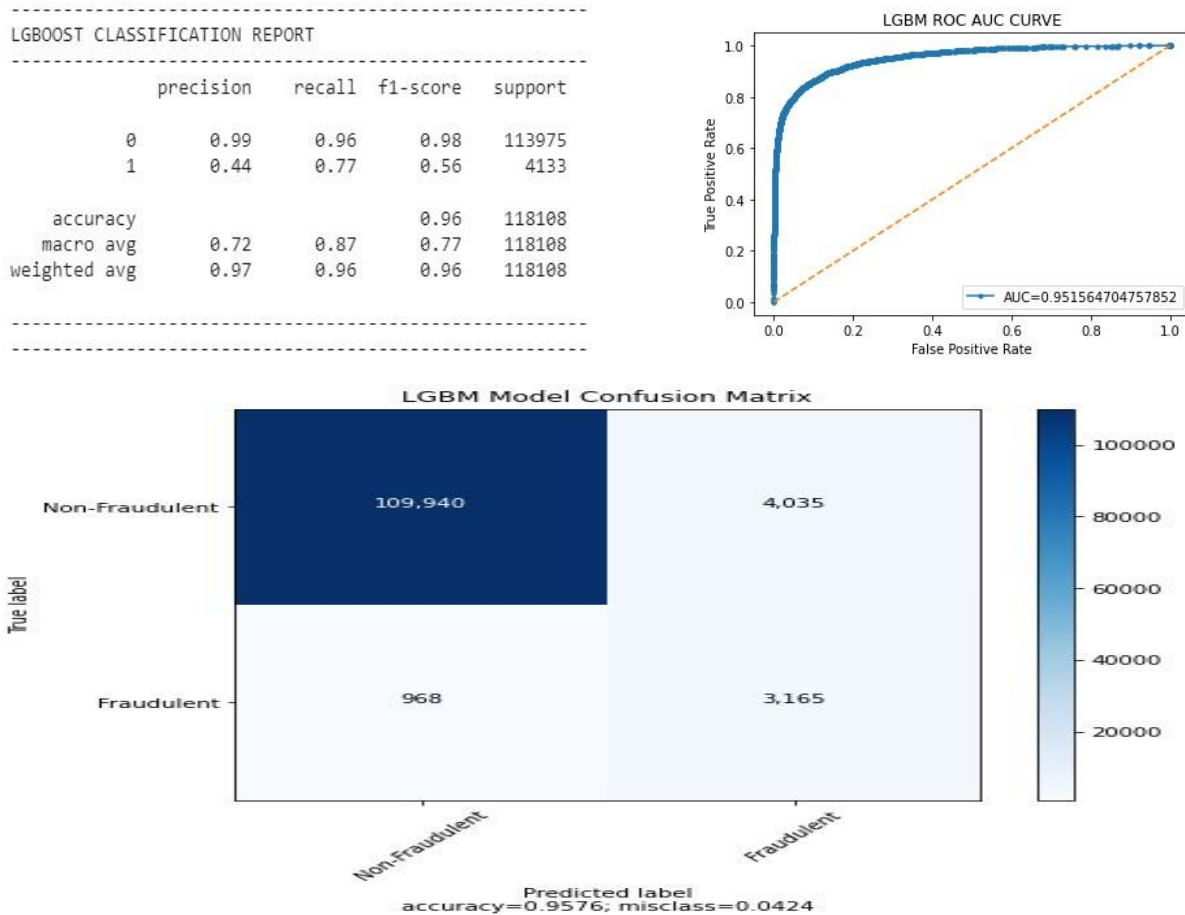


Figure 5.2. LGBM Model Evaluation Report

5.1.3. Random Forest Model Performance

Figure 5.3. below shows the performance of the extreme gradient boosting model on the validation dataset. As shown in the classification report, the model performs better on non-fraudulent events, with an F1 score of 99%. The F1 score on the fraudulent events is 56%. Out of the 4,133 fraudulent transactions, 2,022 were correctly predicted as fraudulent, while the other 2,111 were incorrectly predicted as non-fraudulent. The overall F1 score of the model is approximately 55.8%.

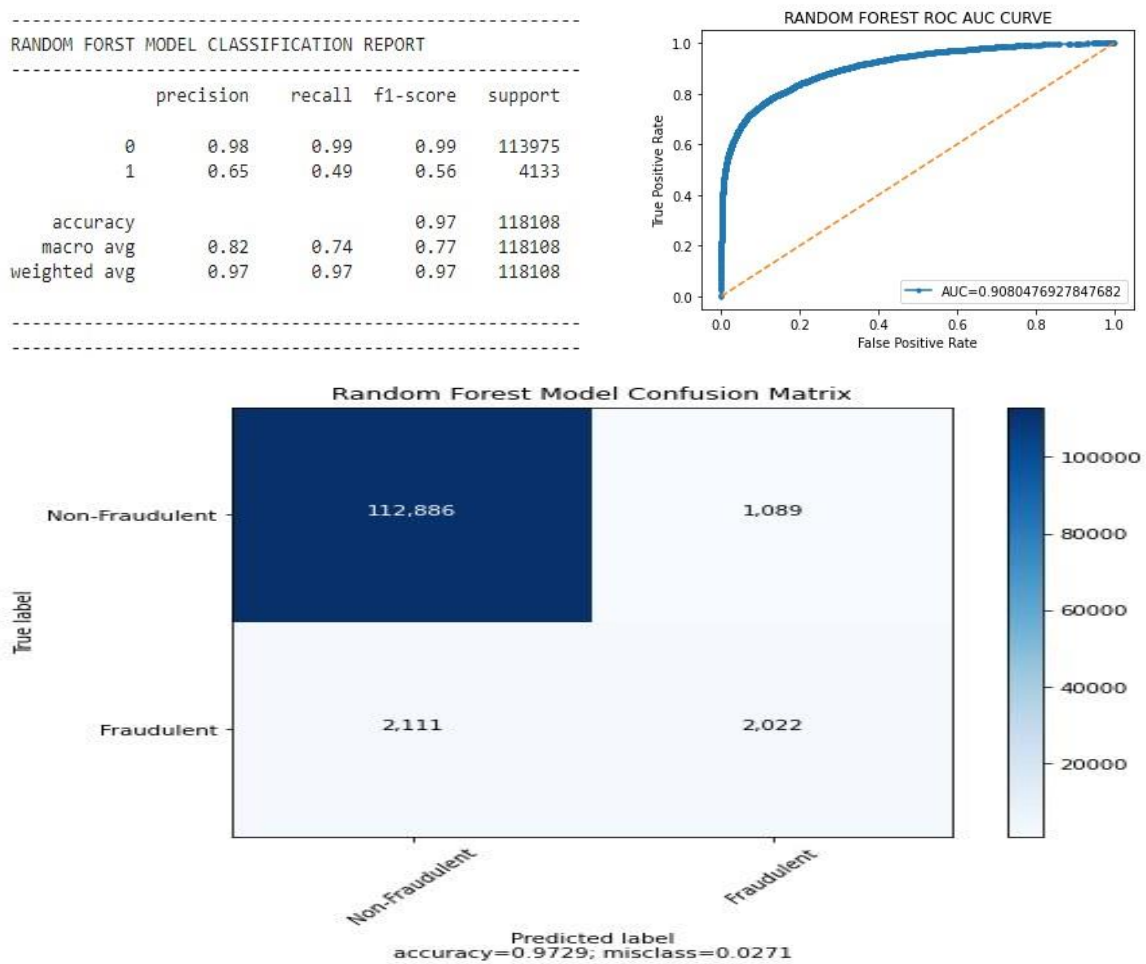


Figure 5.3. Random Forest Model Evaluation Report

5.1.4. Performance Comparison

Model	Accuracy	Precision	Recall	F1
XGBOOST	0.983	0.756	0.761	0.759
LGBM	0.958	0.440	0.766	0.559
SVM	0.966	0.552	0.489	0.505

RANDOM FOREST	0.973	0.650	0.489	0.558
---------------	-------	-------	-------	-------

Table 5.2 Models Performance

Table 5.2 above shows the performance of the ensemble models and the baseline model, Support Vector Machine, with respect to four performance metrics. As shown in the table, XGBoost has the highest accuracy score of 98.3%, followed by Random Forest with 97.3%. Although Support Vector Machine happened to beat the accuracy of the LGBM model by 0.7%, this does not imply that the SVM model is better than the LGBM model because the accuracy score happens to be an unreliable metric in an unbalanced classification problem like fraud detection. The precision score shows that the XGBoost model outperforms the other models with a value of 75.6%, followed by the Random Forest model with a value of 65%. The recall score shows that the LGBM model outperforms the other models with a value of 76.6%, followed by the Random Forest model with a value of 76.1%. The overall performance of the model taken by the F1 score shows that the XGBoost model is the best model for implementation. It also shows that all the ensemble models outperform the baseline model, the Support Vector Machine.

5.2. Model robustness

Different noise proportions were introduced to the training datasets before modelling to reduce the models' generalization error and test their robustness. The robustness of the models is then evaluated by taking note of the effect of the mislabelled data on the models' performance. Since the dataset only contains labels that are either 1s for fraudulent transactions or 0s for non-fraudulent transactions, the noises are introduced by randomly choosing a portion of the dataset and then changing their labels. The models' robustness is assessed with 1% noise addition to the labels. The F1 results of the models against the noise proportion are shown in Table 5.3 below. The table shows that all the model's performance drops with the introduction of noise. At a 1% noise rate, the Random Forest models' performance didn't reduce, while LGBM, XGBoost and SVM models

reduced by 2.9%, 2.4% and 0.4%, respectively. With the introduction of noise, the models' performance only dropped by a small percentage, showing that the models exhibit high robustness.

Model Name	Noise Rate	
	0%	1%
XGBOOST	0.759	0.735
LGBM	0.559	0.530
SVM	0.505	0.501
RANDOM FOREST	0.558	0.558

Table 5.3 Models Robustness

5.3. Conclusion

In the previous chapter, four models—Support Vector Machine, XGBoost Model, Random Forest Model, and LGBM Model—were trained to detect frauds in e-commerce transactions. In this chapter, a thorough model evaluation is done to assess each model's performance. Several tests were run to assess the effectiveness and robustness of the suggested models. The XGBoost model outperforms the other models based on the different evaluation metrics. In addition, the other two ensemble models, Random Forest and LGBM outperform the baseline model. This shows how effective ensembling models are at detecting fraud in e-commerce transactions.

CHAPTER SIX CONCLUSION RECOMMENDATION

6.1. Overview

This research aims to advance e-commerce transaction fraud detection by proposing a novel model with different ensemble machine learning and K-Means features. Various ensemble models were trained and evaluated to demonstrate their viability. The implementation of the ensemble machine learning models showed their advantage over traditional classification models like Support Vector Machine. Different feature engineering techniques were applied to the raw data to improve the performance of the created models. This chapter will wrap up the study and offer a final observation on the problem description and the usefulness of the suggested solution. Additionally, suggestions and potential directions for future studies that can enhance fraud detection are also covered.

Three distinct inference classifiers, Random Forest Classifier, XGBoost, and LGBM, were tested for performance using historical e-commerce transaction datasets. The XGBoost Model, with an F1 score of 75.9%, was the best performing model. In addition to having the minor performance of all the models, the Support Vector Machine model took a very long time of more than 3 hours to train due to the data size.

The result of this study proves that ensemble models, compared to other classification models, are good at detecting fraud in e-commerce transactions, with the XGBoost model coming at the top of all the ensemble models.

6.2. Research Contributions

A list of the contributions made because of this study is provided in this section. The main benefit of this study is the formalization of an ensemble model for fraud detection that e-commerce platforms can apply. The concept's core consists of:

- 1) Feature engineering technique that creates different features to improve the models' performance. The main feature of engineering is the features of an unsupervised machine learning model, K-Means clustering, to learn hidden patterns that classification models do not easily detect.
- 2) Ensembling machine learning models that combine historical transaction data features and the K-Means features to learn patterns that aid them in correctly classifying transactions as either non-fraudulent or fraudulent.
- 3) Model evaluation strategy helps assess the models' capabilities and limitations and defend the conclusions using facts, statistics, and modelling methods. The F1 score is the primary metric used to determine the efficiency of each model.

6.3. Recommendations for Future Work

Future employment options are numerous and diverse. This study's conceptualization of the fraud framework can be applied to various data points from multiple e-commerce channels. Other areas are worth exploring to further improve fraud detection in the ecommerce domain. The domain can be further explored by:

- 1) Implementing Deep Learning Models: Deep learning models like Artificial Neural Networks and Recurrent Neural Networks can be helpful in detecting fraud as these models are good at detecting data sequences.
- 2) More Feature Engineering: More features that can correctly classify transactions as either fraudulent or non-fraudulent transactions are needed to improve the model performance. These features can be created by feature engineering.
- 3) Stacking the ensemble models: Concepts like voting, stacking and blending can be applied to the result of the ensemble models to improve the prediction results. These techniques work by employing several distinct models to produce predictions and then using those predictions as features in a higher-level metamodel. Since each type of lower-level student brings a unique set of strengths to the meta-model, it can significantly improve the prediction result.

REFERENCES

1. Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud Detection System: A Survey." *Journal of network and computer applications* 68 (2016): 90–113. Web.
2. Adnan, Rana Muhammad et al. "Suspended Sediment Modeling Using a Heuristic Regression Method Hybridized with Kmeans Clustering." *Sustainability* (Basel, Switzerland) 13.9 (2021): 4648–. Web.
3. AGARWAL, B. & MITTAL, N. 2012. Hybrid approach for detection of anomaly network traffic using data mining techniques. *Procedia Technology*, 6, 996-1003.
4. AWOYEMI, J.O., ADETUNMBI, A.O. and OLUWADARE, S.A., 2017, October. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNi)* (pp. 1-9). IEEE.
5. BAESENS, B., VAN VLASSELAER, V. & VERBEKE, W. 2015. *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*, John Wiley & Sons.
6. Bai, Xue et al. "Ultrafast Pulse Wave Velocity and Ensemble Learning to Predict Atherosclerosis Risk." *INTERNATIONAL JOURNAL OF CARDIOVASCULAR IMAGING* 38.9 (2022): 1885–1893. Web.
7. Bao, Wang, Ning Lianju, and Kong Yue. "Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment." *Expert systems with applications* 128 (2019): 301–315. Web.
8. BATANI, J. 2017. An adaptive and real-time fraud detection algorithm in online transactions. *Int. J. Comput. Sci. Bus. Inform*, 17, 1-12.
9. Bhattacharyya, Siddhartha et al. "Data Mining for Credit Card Fraud: A Comparative Study." *DECISION SUPPORT SYSTEMS* 50.3 (2011): 602–613. Web.
10. BHATI, P. & SHARMA, M. 2015. Credit card number fraud detection using Kmeans with hidden markov method. *SSRG International Journal of Mobile Computing & Application (SSRG-IJMCA)*—volume, 2.

11. BIAN, Y., CHENG, M., YANG, C., YUAN, Y., LI, Q., ZHAO, J. L. & LIANG, L.
Financial fraud detection: a new ensemble learning approach for imbalanced data.
20th Pacific Asia Conference on Information Systems (PACIS 2016), 2016.
Association for Information Systems, 315.
12. BREIMAN, L. 1996. Bagging predictors. *Machine learning*, 24, 123-140.
13. CHANG, Y.C., LAI, K.T., CHOU, S.C.T. and CHEN, M.S., 2017, July. Mining the networks of telecommunication fraud groups using social network analysis. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 1128-1131).
14. Chang, Hau-Ming et al. "Enhanced Understanding of Osmotic Membrane Bioreactors through Machine Learning Modeling of Water Flux and Salinity." *The Science of the total environment* 838.Pt 1 (2022): 156009–156009. Web.
15. CHEN, J. I.-Z. & LAI, K.-L. 2021. Deep convolution neural network model for creditcard fraud detection and alert. *Journal of Artificial Intelligence*, 3, 101-112.
16. CHOUGULE, P., THAKARE, A., KALE, P., GOLE, M. & NANEKAR, P. 2015.
Genetic K-means algorithm for credit card fraud detection. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 6, 1724-1727.
17. CHOUIEHH, A. and HAJ, E.H.I.E., 2018. Convnets for fraud detection analysis. *Procedia Computer Science*, 127, pp.133-138.
18. Dal Pozzolo, Andrea et al. "Learned Lessons in Credit Card Fraud Detection from a Practitioner Perspective." *Expert systems with applications* 41.10 (2014): 4915–4928. Web.
19. Das, Suchismita et al. "Brain Tumor Segmentation and Overall Survival Period Prediction in Glioblastoma Multiforme Using Radiomic Features." *Concurrency and computation* 34.20 (2022): n. pag. Web.
20. DORNADULA, V. N. & GEETHA, S. 2019. Credit card fraud detection using machine learning algorithms. *Procedia computer science*, 165, 631-641.
21. EOM, W.-J., SONG, Y.-J., PARK, C.-H., KIM, J.-K., KIM, G.-H. & CHO, Y.-Z. Network traffic classification using ensemble learning in software-defined networks. 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), 2021. IEEE, 089-092.

- 22.FU, K., CHENG, D., TU, Y. & ZHANG, L. Credit card fraud detection using convolutional neural networks. International conference on neural information processing, 2016. Springer, 483-490.
- 23.Glas, Afina S. et al. "The Diagnostic Odds Ratio: a Single Indicator of Test Performance." Journal of clinical epidemiology 56.11 (2003): 1129–1135. Web.
- 24.Global Fraud Report 2021, Cybersource and the Merchant Risk Council (MRC), viewed 21 July 2022,
<https://www.cybersource.com/content/dam/documents/campaign/global-fraudreport-2021.pdf>
- 25.GUO, Q., LI, Z., AN, B., HUI, P., HUANG, J., ZHANG, L. & ZHAO, M. Securing the deep fraud detector in large-scale e-commerce platform via adversarial machine learning approach. The World Wide Web Conference, 2019. 616-626.
- 26.Hsiao, H. Y., and K. N. Chiang. "AI-Assisted Reliability Life Prediction Model for Wafer-Level Packaging Using the Random Forest Method." Journal of mechanics 37 (2020): 28–36. Web.
- 27.JHA, S., GUILLEN, M. & WESTLAND, J. C. 2012. Employing transaction aggregation strategy to detect credit card fraud. *Expert systems with applications*, 39, 12650-12657.
- 28.KOTHARI, C. & GARG, G. 2014. Research methodology Methods and Techniques. 2014-New Age International (P) Ltd. *New Delhi*.
- 29.Li, JiaoLong. "E-Commerce Fraud Detection Model by Computer Artificial Intelligence Data Mining." Computational intelligence and neuroscience 2022 (2022): 8783783–9. Web.
- 30.Liang, Minfei et al. "Interpretable Ensemble-Machine-Learning Models for Predicting Creep Behavior of Concrete." Cement & concrete composites 125 (2022): 104295–. Web.
- 31.LIM, H.K., KIM, J.B., HEO J.S., KIM, K., HONG, Y.G. and HAN Y.H., 2019, February. Packet-based network traffic classification using deep learning. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)* (pp. 046-051). IEEE.

32. López, Victoria et al. "Analysis of Preprocessing Vs. Cost-Sensitive Learning for Imbalanced Classification. Open Problems on Intrinsic Data Characteristics." *Expert systems with applications* 39.7 (2012): 6585–6608. Web.
33. Malekipirbazari, Milad, and Vural Aksakalli. "Risk Assessment in Social Lending via Random Forests." *Expert systems with applications* 42.10 (2015): 4621–4631. Web.
34. MAYR, A., BINDER, H., GEFELLER, O. & SCHMID, M. 2014. The evolution of boosting algorithms. *Methods of information in medicine*, 53, 419-427.
35. MYERS, M. D. 2019. *Qualitative research in business and management*, Sage.
36. NANDURI, J., LIU, Y.-W., YANG, K. & JIA, Y. Ecommerce fraud detection through fraud islands and Multi-layer machine learning model. *Future of Information and Communication Conference*, 2020. Springer, 556-570.
37. NAVEEN, P. and DIWAN, B., 2020, October. Relative Analysis of ML Algorithm QDA, LR and SVM for Credit Card Fraud Detection Dataset. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(ISMAC)* (pp. 976-981). IEEE.
38. PAINTAL, S. 2021. ECOMMERCE AND ONLINE SECURITY. *International Journal of Management (IJM)*, 12.
39. PUMSIRIRAT, A. and LIU, Y., 2018. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of advanced computer science and applications*, 9(1).
40. QUEIRÓS, A., FARIA, D. & ALMEIDA, F. 2017. Strengths and limitations of qualitative and quantitative research methods. *European journal of education studies*.
41. RANDHAWA, K., LOO, C.K., SEERA, M., LIM, C.P. and NANDI, A.K., 2018. Credit card fraud detection using AdaBoost and majority voting. *IEEE access*, 6, pp.14277-14284.
42. RASHMI, R., SAMPURNA, J., SHILPA SHREE, N. & SWATI, M. 2018. Credit Card Fraud Detection Using Big Data.
43. Ren, Xudie et al. "A Novel Image Classification Method with CNN-XGBoost Model." *DIGITAL FORENSICS AND WATERMARKING*. Vol. 10431. Cham:

- Springer International Publishing, 2017. 378–390. Web.
44. ROJAS, R. 2013. *Neural networks: a systematic introduction*, Springer Science & Business Media.
 45. RUTBERG, S. & BOUIKIDIS, C. D. 2018. Focusing on the fundamentals: A simplistic differentiation between qualitative and quantitative research. *Nephrology Nursing Journal*, 45, 209-213.
 46. Sahin, Y., Bulkan, S. and Duman, E., 2013. A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, 40(15), pp.5916-5923.
 47. SHONE, N., NGOC, T.N., PHAI, V.D. and SHI, Q., 2018. A deep learning approach to network intrusion detection. *IEEE transactions on emerging topics in computational intelligence*, 2(1), pp.41-50.
 48. SINAGA, K. P. & YANG, M.-S. 2020. Unsupervised K-means clustering algorithm. *IEEE access*, 8, 80716-80727.
 49. SINGH, A. & NARAYAN, D. 2012. A survey on hidden markov model for credit card fraud detection. *International Journal of Engineering and Advanced Technology (IJEAT)*, 1, 49-52.
 50. Su, Qing-Hua, and Kuo-Ning Chiang. "Predicting Wafer-Level Package Reliability Life Using Mixed Supervised and Unsupervised Machine Learning Algorithms." *Materials* 15.11 (2022): 3897–. Web.
 51. Tran, Van Dat. "The Relationship Among Product Risk, Perceived Satisfaction and Purchase Intentions for Online Shopping." *The Journal of Asian finance, economics, and business* 7.6 (2020): 221–231. Web.
 52. VAN VLASSELAER, V., BRAVO, C., CAELEN, O., ELIASSI-RAD, T., AKOGLU, L., SNOECK, M. & BAESSENS, B. 2015. APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, 75, 38-48.
 53. Wang, Gang et al. "Two Credit Scoring Models Based on Dual Strategy Ensemble Trees." *Knowledge-based systems* 26 (2012): 61–68. Web.
 54. WANG, S., LIU, C., GAO, X., QU, H. & XU, W. Session-based fraud detection in online e-commerce transactions using recurrent neural networks. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

Springer, 241-252.

55. WEI, W., LI, J., CAO, L., OU, Y. and CHEN, J., 2013. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4), pp.449-475.
56. WENG, H., LI, Z., JI, S., CHU, C., LU, H., DU, T. & HE, Q. Online e-commerce fraud: a large-scale detection and analysis. 2018 IEEE 34th International Conference on Data Engineering (ICDE), 2018. IEEE, 1435-1440.
57. WENG, H., JI, S., DUAN, F., LI, Z., CHEN, J., HE, Q. & WANG, T. Cats: crossplatform e-commerce fraud detection. 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019. IEEE, 1874-1885.
58. XUAN, S., LIU, G., LI, Z., ZHENG, L., WANG, S. and JIANG, C., 2018, March. Random forest for credit card fraud detection. In *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)* (pp. 1-6). IEEE.
59. ZAMINI, M. and MONTAZER, G., 2018, December. Credit card fraud detection using autoencoder based clustering. In *2018 9th International Symposium on Telecommunications (IST)* (pp. 486-491). IEEE.
60. Zhao, Jie et al. "Extracting and Reasoning About Implicit Behavioral Evidences for Detecting Fraudulent Online Transactions in e-Commerce." *DECISION SUPPORT SYSTEMS* 86 (2016): 109–121. Web.
61. Zhu, Xiaojin, and Andrew B Goldberg. *Introduction to Semi-Supervised Learning*. Vol. 6. San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA): Morgan & Claypool Publishers, 2009. Web.