Credit Card Fraud Detection using Boosted Random Forest Algorithm

By

Thanganedzo Beverly Mashamba
16012810

Supervisor(s):
Prof W Chagwiza
Prof W Garira

Submitted in partial fulfilment of the requirements for the degree of Master of Science Degree in e-Science

in the

Department of Mathematical and Computational Sciences
Faculty of Science, Engineering and Agriculture
University of Venda

Declaration

I, Thanganedzo Beverly Mashamba(16012810), hereby declare that the mini-dissertation titled: " Credit card fraud detection using boosted random forest algorithm " is my own, unaided work. It is being submitted for the degree of Master of Science Degree in e-Science at the University of Venda, Thohoyandou, Limpopo, South Africa. It has not been submitted for any degree or examination at any other university.

Thanganedzo Beverly Mashamba
16012810
29 August 2023

Abstract

Financial fraud is a growing concern with far-reaching concerns in financial institutions, government, and corporate organizations, leading to substantial monetary losses. The primary cause of financial loss is credit card fraud; it affects issuers and clients, which is a significant threat to the business as clients will run to their competitors, wherein they will feel secure. Solving fraud problems is beyond human capability, so financial institutions can utilize machine learning algorithms to detect fraudulent behaviour by learning through credit card transactions. This thesis develops the boosted random forest, integrating an adaptive boosting algorithm into a random forest algorithm, such that the performance of a model is improved in predicting credit card fraudulent transactions. The confusion matrix is used to evaluate the performance of the models, wherein random forest, adaptive boosting and boosted random forest were compared. The results indicated that the boosted random forest outperformed the individual models with an accuracy of 99.9%, which corresponded with the results from confusion matrix. However random forest and adaptive boosting had 100% and 99% respectively, which did not correspond to the results on confusion matrix, meaning the individual models need to be more accurate. Thus, by implementing the proposed approach to a credit card management system, financial loss will be reduced to a greater extent.

Keywords: Adaptive boosting, credit card, evaluation matrices, fraud, random forest.

Acknowledgements

I am profoundly grateful to the Almighty God for bestowing upon me the strength and resilience needed to navigate the challenges of this research project. I extend my sincere appreciation to the National e-Science Postgraduate Teaching and Training Platform (NEPTTP) for their generous sponsorship, which made this endeavour possible. I am indebted to my dedicated supervisors, Prof. W Chagwiza and Prof. W Garira, for their invaluable guidance and mentorship throughout this research. Lastly, I extend my heartfelt thanks to my unwaveringly supportive family and friends, whose encouragement and belief in my abilities have been a constant source of motivation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of the study

Financial frauds is a growing concern with far-reaching concern in the financial institution, government, cooperate organizations and citizens in general. There is high internet technology in today's world, which leads to an increase in credit card fraud being the most concerning problem around the world. There are many fraud detection solutions and software which detect fraudulent activities in organizations like banks, retails, e-commerce, insurance, and other industries, to name a few. Machine learning techniques are the most notable and popular methods used in solving credit card fraud.

Many people have started believing in going cashless whenever they make their purchases. A credit card has made it easier for online transactions and accessibility. A credit card allows someone to borrow money from the bank against a line of credit. The card is then helpful to make purchases and pay for other expenses. If not used responsibly, this credit card can end up being in the wrong hands, which can lead to fraud, which is why credit card fraud has taken a lot of companies attention, especially the ones offering credit cards.

Credit card fraud is an inclusive term for fraud committed using a payment card, such as a credit card [1]. The purpose may be to obtain goods or services or make payment to another account that a criminal controls. According to [79], in South Africa, between 2018 and 2019, credit and debit card fraud increased by 20.5% due to the increase in making online transactions, and from their analysis, 66.6% of frauds were made from foreign countries, which shows that fraudulent activity

are not limited to geographical area.Therefore Credit card fraud is a problem that needs to be addressed or evaluated more often to ensure that the techniques in place are still up to standards, including introducing new and improved approaches that will help in detecting the credit card fraudulent transactions from both local and foreign countries.

According to [13], along with the evolution of fraud detection methods, perpetrators of fraud have also been evolving their fraud practices to avoid detection, and [14] claims that the method used in credit card fraud detection need constant evaluation. The credit card fraud problem does not only affect credit card issuers, but also customers [29]. According to [17], customers are often refunded by banks whenever they are victimized by fraud. However, [60] claims that as a result, customers suffer emotionally, wherein some end up hospitalized and, worst part committing suicide. As we can see, this will harm both credit card issuers and customers. According to [12], credit card issuers are the most affected party in credit card fraud as they have to accept entire liability for losses due to fraud, which can affect an organization's future performance. Without a doubt, therefore, credit card fraud is a recurring problem for all credit card issuers. In this study, we evaluate a supervised machine learning algorithm, random forest, which will be integrated with a boosting algorithm, AdaBoost, as part of an attempt to detect credit card fraud better.

Machine learning models for detecting credit card fraud are in use. Despite the proliferation of data mining techniques and applications in recent years, few published studies of data mining have been published for credit card fraud detection. The majority of these papers have explored neural networks [22, 25, 26, 49, 55], which is not surprising considering their success in the early years 1990s till today. Some papers including recent ones [21, 34, 45, 61, 72, 78] evaluates several techniques, including support vector machine, neural network and random forest for detecting credit card frauds. Some of these papers focus on the impact the effect of aggregating transactional data on fraud detection results; results were investigated on aggregation over various time periods on two real-world datasets and discovered that aggregation could be beneficial, with aggregation duration length being a significant factor. Aggregation was discovered to be particularly successful with random forests.

Random forests are advanced and most popular machine learning technique that has been shown in recent years to outperform in a variety of applications [11, 75, 64, 86, 36]. The random forest and AdaBoost for boosting random forest were chosen for this study due to its accessibility for practitioners, ease of use, and noted performance advantages in the literature. There are no known studies that deal with credit card fraud detection using a boosted random forest algorithm, but papers show the significance of using boosted random forest algorithm [50, 80, 38, 32]. Below is the conceptual framework of the credit card fraud detection process:

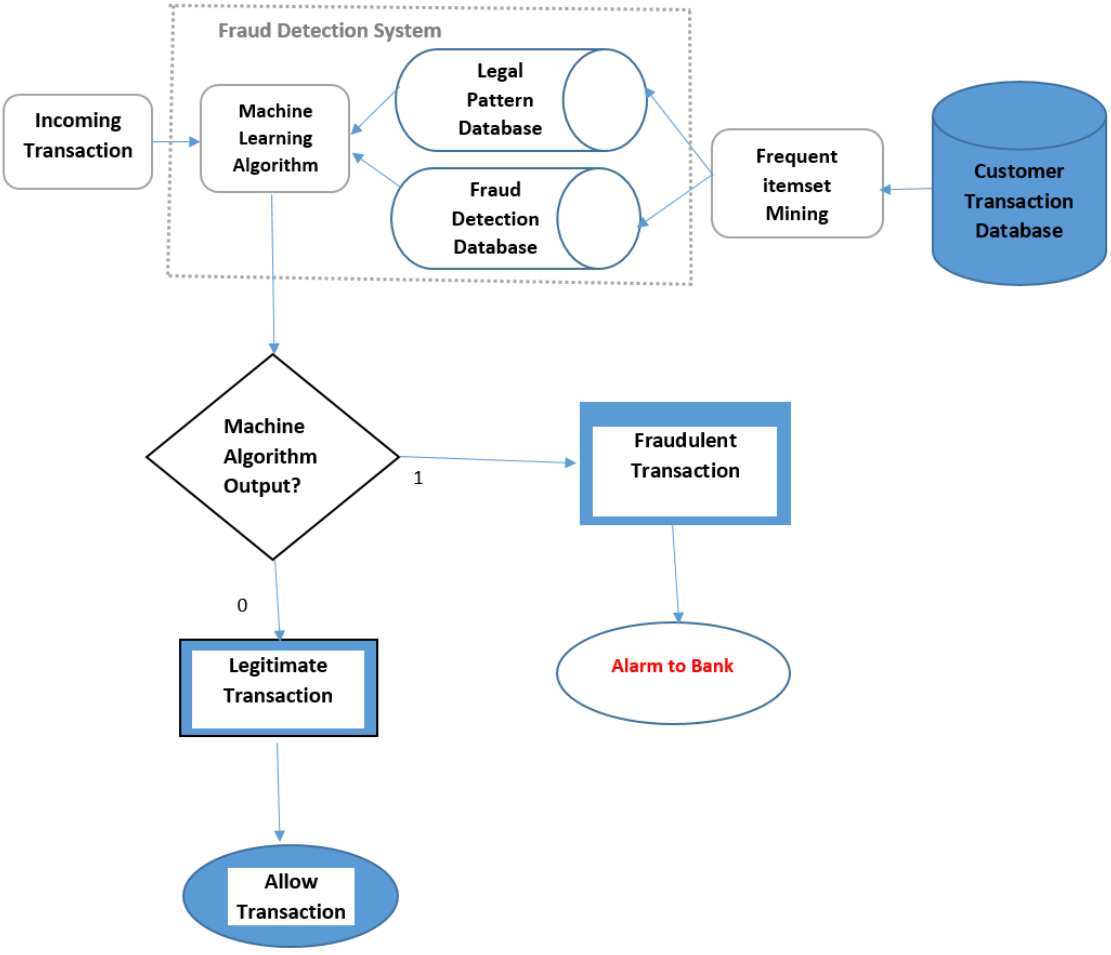


FIGURE 1.1: Conceptual framework for credit card fraud detection

Boosted random forest algorithm, can be embedded on the credit card system responsible for monitoring credit card transactions. Boosting is a widely known ensemble training algorithm used to build random forest classifiers in a sequential manner using independent decision trees. Boosting is an ensemble training algorithm that combines weak learners with low discriminating performance individually to create a higher discriminative performance classifier. Boosting achieves high discrimination by sequentially training classifiers, with the previous classifier's classification errors used to train the next classifier to generate accurate classification. On the other hand, Boosting tends to over-fit the training sample, whereas random forest avoids overfitting to the training sample by using randomness in the decision tree construction. As a result, a large number of decision trees must be built in order to achieve high generality.

This work evaluates the machine learning algorithm approach, which is boosted random forest algorithm. The performance of a model will be evaluated using a confusion matrix and the model's accuracy. Model's performance will be based on the credit card transaction dataset, and this dataset is highly imbalanced, skewed, and rarely available. Feature selection, feature engineering, and suitable evaluation matrix is the most crucial part of a machine learning algorithms to evaluate the performance of a model on a high imbalance and skewed credit card transaction dataset.

1.2 Problem statement

Fraud detection in credit card systems is the most concerning issue. With the growth of information technology and improved data communication, fraud is changing since fraudsters finds new ways to prevent detection and commit theft, causing substantial financial losses and harm to customers which might results in customer loss. So it is crucial for credit card issuers to make ensure that the system in place for detection of fraudulent activities are still up to standard through constant evaluation, this will improve the detection hence preventing such behaviours from happening even though eliminating it is impossible. Fraud detection and prevention are beyond human capability, so using machine learning techniques provide promising results with more probability of detecting fraudulent behaviour. The

study proposes the usage of a boosted random forest algorithm for detection of credit card frauds.

1.3 Research Questions

The following are research questions that will be answered in this study:

a) How customer's behaviour impact credit card fraud detection?

b) What impact does feature engineering have on the performance of a boosted random forest approach?

c) What impact does feature selection have on the performance of a boosted random forest approach?

d) How accurate can a well-trained boosted random forest detect credit card frauds in the banking industry?

1.4 Research Aim and Objectives

This section outlines the aims and objectives that will guide answering the research questions.

1.4.1 Research Aim

This study aims to develop boosted random forest model that will detect fraudulent activities in transactions made using credit cards.

1.4.2 Objectives

The objective of the research is:

a) To extract the behavioural pattern of customers through analysing the dataset;

b) To create appropriate features for the boosted random forest model;

c) To develop a comprehensive feature selection process using random forest model; and

d) To design and develop a boosted random forest algorithm for credit card fraud detection.

1.5 Scientific Contribution

There are a lot of predictive models thus far used for predicting credit card frauds. Since credit card fraud tends to cause a massive loss in financial institutions, it is crucial for credit card issuers to pay much attention to fraud issues and use machine learning algorithms to predict such behaviours before credit card transactions are approved. It is essential for credit card issuers to have a model that will be accurate enough to detect fraudulent behaviours. There are success stories that showed great performance of the boosted models applied in areas like healthcare and agriculture [3, 31, 33, 39]. The study propose the use of boosted random forest algorithm for credit card fraud detection. Random forest is known to be the extension of bagging algorithm used on weak learners which exhibit high variance and low bias. Random forest is integrated with a boosting algorithm, adaptive boosting algorithm, which is leveraged when variance low and high bias is observed. Therefore combining the bagging and boosting algorithm will be the model with low bias and high variance resulting in an optimal boosted random forest model.

1.6 Limitations

This study has a potential limitation. The dataset used in the study will be secondary data, We were hoping to find some recent data of people in South Africa or Africa in general, but there was no open data; this means that the current credit card fraud detection will be based on the old credit card transaction data.

1.7 Overview

The research proposal is structured as follows; Chapter 1 introduces the reader to the study area, defines the problem to be studied, and proposed a solution. Chapter

2 discusses the existing literature within the research area. Chapter 3 presents a machine-learning algorithm to be used in the study and outlines the experimental design in terms of the dataset, data pre-processing, and data exploratory analysis, including data visualization and choices of evaluation matrices. Chapter 4 discuss the results obtained from the stud carried out to answer some questions. Chapter 5 presents the working schedule and potential difficulties. Chapter 5 conclude the study and also make some recommendation on the feature work.

Chapter 2

Literature Review

2.1 Introduction

The study investigates the performance of the boosted random forest in the detection of credit card fraud. Detecting credit card frauds with machine learning algorithms is a common topic among researchers. However, no evidence showed fraudulent detection using the boosted random forest, but the boosted random forest model has shown outstanding performance when applied in different applications. Credit card fraud is a concerning problem worldwide wherein most parts of the continent are affected more than the other. Additionally, this is because the lifestyle, environment, race, and culture differs with each group of people; hence, this section will be reviewing various papers that analysed various methods on a different dataset from different places to understand how each model was performing. Therefore, this section briefly presents some of the techniques used to detect credit card frauds previously and the methods used in feature engineering and selection.

2.2 An overview of credit card fraud around the world

Credit card fraudulent behaviour is one of the biggest threats to both the business and customers globally, and it has increased drastically. According to [44] some countries are highly targeted by the fraudsters regardless of having improved technologies, with the United States (US) being one of the most targets around the

world, with Covid-19 playing a vital role in the growth of credit card frauds. According to [63], card payment fraud loss was approximately 28,56 billion dollars worldwide, with the United States being responsible for more than a third of the total global loss. Additionally, it resulted in it being the most credit card fraud-prone country in the world. Credit card fraud has also increased in the Southern part of Africa due to the Covid-19 pandemic. Furthermore, changes like human behaviour, human movement, and policing, creating new platforms for criminals was triggered by covid19 and crucially impacted the number of crime incidents. However, it does not matter the type of improved technology used since the same improved technology used in the credit card system is the same technology used by fraudsters to try and find ways to prevent and stop fraud detections. Credit card issuers must constantly innovate the technology used in credit card management systems.

2.3 Understanding the credit card customer's spending patterns

In the detection of credit cards frauds is crucial for credit card issuers to understand their customer's spending behaviour. Because how customers use their credit cards differs with things like age, gender, race, education, job and the environment they live in, to name a few. Moreover, from that personal information, as credit card issuers, we can discover the patterns that fraudsters can use to target a particular customer. Hence it can assist credit card issuers in finding a way to prevent fraudulent transactions. Spending patterns refers to why and how behind customer purchasing decisions, indicating all the habits and routines that the customers show through the products and services they purchase. Moreover, understanding customers' spending behaviour is crucial for the detection process. It also allows the credit card issuer to know their customer's preferences and factors influencing customers purchase decisions [73]. Moreover, this will enable issuers to understand things like where the customers work, where they stay, what they usually buy the most, products they are interested in, what drives them into deciding on purchasing something.

There are tools and methods in places that business retailers have used to understand the customers' spending behaviour. Most small business retailers are still using traditional methods like Google Analytics, Facebook insights audience, HubSpot Service Hub to understand customers' behaviours. At the same time, large businesses use more advanced tools like machine learning techniques to analyse their customer's behaviour. Moreover, it has proved to be ideal. It looks at the fact that machines can use the same information in helping monitor all transactions, including detecting some abnormal transactions. This study focuses on the machine learning tools used to understand the customer's behaviour.

The study [2]carried out some investigation using a questionnaire survey to understand the customer's attitude and spending pattern with credit cards. The test was performed with attitude being the dependent variable and the independent variables being: self-esteem, lifestyle, peer group pressure, time consciousness, and exposure to gregariousness and advertisement. The predictions were made using a regression model to check the relations between the dependent and independent variables and discovered that only customers' lifestyles and attitudes were significant influencers of credit card customers' spending behaviour. Therefore, lifestyle and attitude were the foremost influencers for people staying in Malaysia.

Some studies [56] looked into the unique features of the customer wherein they analysed the socio-economic, demographic and banking-specific deferments which are influencing the selection of the credit card. The multinomial logit model was used on banking information for customers in Italy, where the type of credit card was used by the model as the dependent variable and set of exploratory variables. The results indicated that women, older people, people residing in the centre of Italy and secondary credit card owners are likely to get a classic card. Moreover, credit card issuers can quickly develop a simple detection strategy by understanding their customer behaviour with a credit card.

Therefore, the study mentioned above indicates that customers' behaviour with credit cards differs with a specific group of people, age, and environment. Hence, all credit card issuers need to understand the customers' behaviour to identify

fraudulent behaviours within credit card transactions.

2.4 Feature engineering methods

In machine learning studies, we often use the raw data, which also contains different features that are important and not relevant to the study, which is why feature engineering is needed on the dataset before adding it to the model. Moreover, feature engineering improves the predictive model's performance [73], which also adds to why it is crucial because it allows the creation of features that are important towards the study's main objectives, hence compatible with the model. Feature engineering is the most crucial part of machine learning. Feature engineering is creating new features from existing features on the dataset following the target to be learned and looking at machine learning algorithms [40]. Additionally, this process includes the transformation of data towards the forms that better relate to the underlying target to learning; hence it increases the effectiveness of the model [51]. As noted in [8] feature engineering is a vital step in machine learning because it allows designing artificial features into an algorithm. That algorithm then employs these artificial features to enhance its performance or, in other words, reap better results [84].

2.4.1 Feature engineering techniques for machine learning

Some data are made from datasets taken from different databases containing both structured and unstructured datasets where each information will have its strength and weakness. Therefore this implies that before feeding the data to a machine learning model, it is vital to transform the original dataset into exciting features that improve the performance of predictive models. The study [7] created new features with the use of Recency, Frequency, Monetary (RFM) principle, wherein researchers analysed the fraudulent transactional dataset preparing it for the predictive models-namely, logistic regression and classification trees; with the engineered data, there was an improvement in the credit card predictive model.

Based on data normalisation and anomalies, data mining algorithms were presented to learn the structure and features of fraudulent and legitimate transactions. [81]. Bayesian network classifiers, specifically K2, nave bayes, tree augmented nave bayes (TAN), J48 classifiers, and logistics, were proposed to predict credit card fraud. The dataset was pre-processed using normalisation methods and principal component analysis, and as a consequence, all classifiers achieved an accuracy of greater than 95%.

In detecting credit card fraud, the random forest was the proposed method, which was analysed on a real-world dataset containing fraudulent transactions [42]. The existing feature from the dataset was engineered using data normalisation to form new exciting features, which was fed to the random forest model. As a result, the model achieved 90% accuracy due to the features obtained from normalised data.

The Deep Convolution Neural Network (DCNN) was a proposed model for credit card fraud detection [18]. The analysis was based on the real-world financial data, wherein the dataset was re-scaled between 1-10, and the process of re-scaling the data is known as data normalization. As a result, after 45 seconds, the DCNN model achieved a detection accuracy of 99%.

In constructing a fraud detection model, selecting essential features from the dataset is crucial, and this is usually done by aggregating the credit card transactions to analyze the customer's purchases behaviour. The study [9] expanded the transaction aggregating strategy by proposing a set of new features which is based on analysis of the periodic behaviour of transactions with the use of Von Mises distribution. The dataset used in the study was a real dataset containing credit card frauds obtained from a large European card processing company. They evaluated different datasets of features that impact results and discovered that the proposed periodic feature engineering methods show an average savings increase of 13%.

The study [83] offered an interesting feature engineering methodology for using deep learning models to construct the detection systems for credit card fraudulent transactions. A feature creation methodology based on homogeneity-oriented behaviour analysis (HOBA) was provided to construct the fraud detection model.

Then, learning algorithms were implemented into the system used for fraud detection to achieve good detection performance. The study was based on real-world data from China's top commercial banks. The experimental findings show that the proposed methodology is a viable mechanism for detecting credit card fraud. For future work, they want to work on the computational requirements of a real-time system for fraud detection and look at how improved machine learning approaches and various integrations of deep learning and standard data mining methods may be used to detect fraud.

In order to solve the credit card fraud problem on the UCSD-FICO data containing the fraudulent transaction, the supervised learning models were proposed; namely, Random forest, decision trees and gradient boost trees [35]. The original dataset was normalised to ensure that the dataset was between the same ranges since other records had higher values than others. The proposed model showed some improvement in the performance after new data was engineered.

The [65] carried out analysis by comparing the predictive power's algorithmic impact spanning three different supervised classification models: logistic regression, gradient boosted trees and deep learning. The [65] also investigated the advantages of creating features with domain expertise and feature engineering with an auto-encoder, an unsupervised feature engineering technique. Furthermore, based on their findings, the researchers concluded that feature engineering with domain expertise significantly improves predictive power. Furthermore, the auto-encoder allows one to lower the dimensionality of the data while somewhat increasing predictive ability.

The study focused on automated feature engineering using hidden Markov model (HMM) to detect credit card fraud [48]. The hidden Markov represent was combined with a cutting-edge expert-based feature engineering technique to model temporal correlations to improve classification job performance and detect fraudulent transactions. The findings suggested that the HMM-based feature engineering technique is essential with promising fraud detection capabilities. Similarly,

HMM-based features could be developed in any supervised job that involves a sequential dataset. According to the reviews on this study on the feature engineering approaches, most studies employed normalising data techniques to build new datasets within the same range, e.g. between 0 and 1; this technique showed to be enhancing the performance of predictive/detection models selected in credit card fraud problem; hence data normalization will be used in this study.

Feature engineering is necessary for all problems that are solved using machine learning models; this is because it allows the creation of new features from the existing features that are compatible with the model hence increasing the model's efficiency and effectiveness.

2.5 Feature selection methods

Following feature engineering is the feature selection process wherein out of the feature created during data pre-processing, we need to select only feature that are important to the study and compatible with the selected machine learning models. Furthermore, this increases the model's performance since the number features would have been reduced.Feature selection, the process of extracting the most consistent, non-redundant, and relevant features in model creation, is known as feature selection. As the number and variety of datasets grow, it is more critical than ever to reduce them methodically. The fundamental purpose of feature selection is to improve the predictive model's performance while lowering the modelling cost. Feature selection strategies reduce the number of input variables by removing redundant or unnecessary features and restricting the set of features down to the ones most useful to the machine learning model.

2.5.1 Feature selection techniques for machine learning

The random forest is a supervised machine learning algorithm used as a feature importance selector. The study [28], in intrusion detection, used the random forest to score features according to how important the feature was towards the study. The

data used in the paper was DAPRA 98 dataset, which provided the labelled data for intrusion detection. The experimental results indicated that the proposed random forest-based approach could choose the most essential and relevant features useful for classification, reducing the number of input features and time and increasing classification accuracy.

In the prediction of prostate cancer, the random forest was used for feature selection [30], which was meant to increase the predictive model's performance. The method used for selecting essential features to the study increased the model accuracy up to 87%. Therefore, the study concluded that random forest is suitable for selecting important features.

[47] conducted research on sequence-based prediction of DNA-binding proteins based on hybrid feature selection utilizing random forest and Gaussian naive Bayes. The proposed feature selection approach combined random forest and wrapper-based feature selection with the forward best-first search strategy to rank the features. DBPred is another proposed technique that uses Gaussian naive Bayes as the underlying classifier. It outperformed five other classifiers: decision tree, logistic regression, k-nearest neighbour, support vector machine with the polynomial kernel, and support vector machine with radial basis function. Because of the random forest feature, the proposed approach outperformed all other classifiers. As a result, all of the experimental results imply that the proposed DBPPred can be employed as an alternative perspective predictor for the large-scale identification of DNA-binding proteins.

The practical feature selection study was carried out because feature selection is vital in speeding up the learning and improving the quality of the concept. The study proposed a new algorithm for feature selection called relief; the algorithm uses statistical methods and avoids heuristic searching [37]. Even though the approach does not always locate the small subset of features wherein the size is usually minimal because only statistically significant features were chosen. [37] focused on the test findings found in two artificial domains, the LED display domain and the parity domain with and without noise. Relief beats other feature selection methods regarding learning time and concept accuracy, emphasizing Relief's value.

[85] suggested a random forest-based feature selection approach for producing low-cost feature subsets by including feature cost into the construction of a basic decision tree process. A feature was chosen randomly when building a base tree, with the probability proportional to its associated cost. The random forest method is tested on the UCI dataset, where it is used for a medical diagnosis problem where the fundamental characteristic is estimated. The results showed that their feature-cost-sensitive random forest (FCS-RF) could select a low-cost subset of relevant features and outperform other state-of-the-art feature selection approaches in real-world applications.

The study [43]proposed using the Boruta R package to create a novel feature selection method for locating all features relevant to the study. The algorithm was built as a wrapper for the random forest classifier. The proposed strategy eliminated those aspects that statistical tests demonstrated to be less meaningful than random probes. The Boruta package provided an easy way to interact with the algorithm.

The study [68] used the random forest to evaluate the feature selection and classification of leukocytes. Many artefacts/noise are extracted during the automatic segmentation of leukocytes from the intricate morphological background of tissue slice pictures. Many artefacts/noise are extracted, resulting in a considerable amount of multivariate data creation. This multivariate data impairs a classifier's capacity to distinguish between leukocytes and artefacts/noise. However, when compared to a high-dimensional features space, the selection of crucial characteristics played a significant influence in reducing computing cost and improving classifier performance. As a result, the Gini importance-based binary random forest feature selection approach is introduced in this study. According to the results, the suggested method successfully reduces redundant information while keeping good classification accuracy compared to earlier feature reduction methods.

The study [71] explored feature selection-based adaptive credit card fraud detection techniques J48 decision tree, AdaBoost, Random Forest, Naive Bayes, and PART were employed in this study to detect financial fraud on credit cards. Their performance was compared using five parameters: sensitivity, specificity, precision, recall,

MCC, and accuracy. The tests were conducted using German credit card data, and the efficiency of the detection mentioned above strategies based on filter and wrapper feature selection methods was assessed. The results showed that utilising the filter and wrapper approaches enhanced J48 and PART prediction accuracy. Finally, the precision and sensitivity of J48, AdaBoost, and the random forest increased.

A various studies on feature selection have been reviewed, with a lot of them focusing on using the random forest to select important features automatically. Therefore, this study used a slightly different approach wherein features are selected manually after the random forest has ranked them; only feature with the highest scores will be selected.

2.6 Machine learning algorithms used in the detection of credit card frauds

The paper [27] examined neural networks' performance in detecting credit card fraud. Over two months, the neural network-based fraud detection system was trained and tested on account transactions that included activity from lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud, and non-received issue fraud. The study employs supervised learning. The network detected more fraudulent accounts with fewer false positives than rule-based fraud detection. The networks' performance was then evaluated based on the accuracy and timeliness of detecting a fraud. The system was installed on an IBM 3090 at Mellon Bank and used for fraud detection on that bank's credit card portfolio.

The papers [58, 74] investigated machine learning algorithms on credit card fraud detection, where evaluations were made on a real-world dataset. However, the approach from both studies is different. Since study [74] evaluated several machine learning algorithms as well as meta-learning, and Tested techniques are ID3, CART, BAYES, and RIPPER. Whereas study [19] evaluated the performance of computational intelligence based on clustering and filtering capabilities of self-organizing maps. Clustering helps in identifying new hidden patterns in input data, which

cannot be done using statistical methods, which is why the researchers [19] concluded that the approach used achieved efficient and cost-effective accurate analysis. The results from [74] indicate that the RIPPER and CART had the highest True Positive rate (80%), but a lower False Positive rate (16%), and the results were equivalent. ID3 works reasonably well, but BAYES outperforms the others by classifying nearly every transaction as non-fraud. According to researchers using [74], a 50%/50% distribution of fraud/non-fraud training data produced classifiers with a high actual positive rate and a low false-positive rate. The best approach discovered was meta-learning with BAYES as a meta-learner to combine base classifiers with the highest True Positive rates learnt from a 50%/50% fraud distribution.

For credit card fraud detection, the study [4], Used a database-driven mining strategy. The technology employs a neural network method, and CARDWATCH includes the neural network technique. The system was tested on synthetically created data using an auto associator. The acquired results revealed that successful fraud detection rates with an accuracy of 85% and legal transaction identification rates of 100% were achieved.

The study [15] investigated credit card fraud detection. The neural network was selected for the study. However, it was then combined with other data mining techniques due to some obstacles of using neural networks alone. Results proved the combination of the techniques to successfully obtain a high fraud coverage combined with a low false rate alarm.

The paper [23] discussed the mechanical design of user profiling for fraud detection using a range of data mining approaches. In a massive database of consumer transactions, rule-learning software was employed to find signs of fraudulent behaviour. The indicators were then utilized to build a set of monitors that monitored normal client behaviour and alerted to anomalies. The monitor outputs were used to combine evidence to generate high-confidence alarms. The method has been used to detect cellular cloning fraud based on a call records database. The results show that the intuitive technique outperforms hand-crafted strategies for detecting fraud.

Studies [19, 70], evaluated the performance of classification models in detecting

credit card frauds. [70] evaluated the performance of decision trees, neural network, and logistic regression. The results from [19, 70] indicate that neural network and logistic regression outperforms decision trees. However, study [19] evaluated the support vector machine (SVM) and Artificial neural network. The discovery is that SVM and ANN are comparable in training, but ANN tends to overfit training data. Hence it will have poor performance in feature predictions if the dataset is small.

The study [66] proposes sporadic principles for detecting credit card fraud because occasional rules are the most well-studied models for data mining. Association rules are utilized to extract information, resulting in standard behaviour patterns in irregular transactions, ensuring fraud detection and prevention. The data used comes from Chilean retail companies. Association rules help overcome challenges with less support and confidence, minimise execution time, decrease superfluous rule development, and contribute more intuitive results.

[53] suggests using Dempster-Shafer theory and Bayesian learning to detect credit card fraud. The detection system comprises four parts: a rule-based filter, a Dempster-Shafer adder, a transaction history database, and Bayesian learning. Based on the first assumptions, the transactions were normal, abnormal, or suspicious. Bayesian learning was utilised to reinforce or weaken the belief based on the resemblance with the fraudulent transaction history. As a result of rigorous modelling using stochastic models, a fusion of other evidence has a powerful beneficial influence on the performance of credit card fraud detection systems.

The papers [6, 41] carried out a research on credit card fraud detection in real-world scenarios using the random forest model to detect frauds in credit card transactions. However, the models were evaluated using a different evaluation matrix. Wherein [6] evaluated the model's performance based on the accuracy, sensitivity, specificity, and precision. The results indicated the random forest model's best performance with 98.6% accuracy, whereas the study [41] evaluated the model's performance based on accuracy and confusion matrix. Their results indicated that the optimal accuracy for Random Forest is 90%.

The study [20] undertook an evaluation of strategies utilised in credit card fraud detection; their key goals were first to identify the various types of credit card fraud and, second, to analyse alternative fraud detection approaches. The secondary goal was to present, compare, and analyse newly published credit card fraud detection findings. [20] covered how to detect various sorts of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud, and behavioural fraud, as well as detection methods such as pair-wise matching, decision trees, clustering approaches, neural networks, and evolutionary algorithms.

The paper [5] identified fraud detection as one of the most common difficulties in the secure banking research sector for reducing the losses of banks and e-transactions organizations; wherein logistic regression (LR), random forest (RF) and modern classifiers such as XGBoost (XG) and CatBoost (CB) were used to test the effect of unbalanced data. Furthermore, by comparing the results with and without balancing and then focusing on the savings measure to test the impact of cost-sensitive wrapping of Bayes minimum risk (BMR), they used F1-score, AUC, and savings measures. The results show that CB has the best savings with 71%, 97% when using SMOTE, and 98 per cent when combining SMOTE and BMR, while XG has the best protection of 76% when using BMR without SMOTE.

The paper [69] is on the prediction of credit card fraud; the researcher explored and tested the effectiveness of Random Forest, AdaBoost, XGBoost, and LightGBM Classifiers on a heavily distorted credit card fraud dataset. The model's performance was assessed based on its accuracy, sensitivity, specificity, and precision. The results showed that Random Forest, AdaBoost, XGBoost, and LightGBM classifiers had the best accuracy 85%, 83%, 97.4%, and 93%, respectively.

The paper [62], researched the detection of credit card fraud using a genetic algorithm, an optimisation technique, and an evolutionary search based on genetic and natural selection principles, a heuristic used to solve high complexity computational problems.

The study [52] presented methods they used in the detection of credit card frauds,

and those methods are decision tree, genetic algorithm, meta-learning strategy, neural network, and HMM. In contemplate system for fraudulent detection, the artificial intelligence concept of support vector machine (SVM) and decision tree were used to solve the problem. Thus, by implementing their approach, financial losses could be reduced to a greater extent.

The paper [76] discussed automated credit card fraud detection using a machine learning algorithm. They applied two machine learning algorithms: artificial neural and Bayesian belief networks. Their results concluded that both techniques have promising results even in the feature.

The project [77] mainly focused on credit card fraud detection in the real world by applying a random forest algorithm; they evaluated the model's performance based on accuracy, sensitivity, specificity, and precision. Their results indicated that the optimal accuracy for Random Forest is 98.6%.

The study [59] focused on real-time fraud detection and proposed a new innovative approach towards understanding how customers spend their money to decipher potential fraud cases. The study made use of the self-organising map to decipher, filter and analyse customer's behaviour for fraud detection.

The study [67] used the association rule since they are considered best-studied models for data mining. The proposed methods extract knowledge so that standard behaviour patterns may be obtained in unlawful transactions from transactional credit card databases for fraud detection and prevention. The analysis was done on the data about credit card fraud in some of the most important retail companies in Chile.

The study [54] suggested an innovative method for detecting credit card fraud by combining current and past behaviour evidence. The fraud detection system comprises four parts: a rule-based filter, a Dempster–Shafer adder, a transaction history database, and a Bayesian learner. The rule-based component assessed the level of suspicion for each incoming transaction based on the extent of its divergence from a good pattern. The Dempster-theory Shafer's was used to combine many sources

of evidence to compute shreds of evidence and an initial belief. The transactions were categorized as normal, abnormal, or suspicious, depending on the first belief. Once a transaction is suspected, belief is raised or lessened based on its closeness to fraudulent or authentic transaction history, as determined by Bayesian learning. The extensive simulation with stochastic models demonstrates that the fusion of multiple pieces of data has a significantly positive impact on the performance of a credit card fraud detection system compared to other methods.

The paper [24] proposed an ensemble model based on sequential modelling of data with a deep recurrent neural network to detect credit card fraudulent activities. Furthermore, as mentioned earlier, they presented a novel algorithm for training the voting. The analysis was based on two real-world datasets; the proposed model outperformed the state-of-art models in all the evaluations. Finally, the time analysis of the proposed solution is efficient and effective in terms of real-time performance versus the recent models in the field.

2.7 Summary

The majority of the studies evaluated the performance of various stand-alone machine learning models applied on a different dataset, and the models were performing well. Most of the studies evaluated their model's performance based on the most used evaluation matrix for classification models and by looking at the model's accuracy. Some of the studies used a traditional statistical model for fraud detection. Overall, the results from the studies show that a model's performance in the detection of credit card frauds varies in terms of the dataset being used in the studies, models being used in the study, processes involved in training the model, and matrices used in evaluating the model's performance. Therefore, in this study, we introduce the combination of boosted random forest model, which is made up of combine two models, namely, Adaptive boosting and random forest; the model's performance will be evaluated using a confusion matrix and by looking at the model's accuracy.

Chapter 3

Research Methodology

3.1 Introduction

This chapter gives a brief account of the methods applied in the research during the analysis and training of the predictive model. The analysis in the study is based on secondary data. Before feeding the data to the model, various processes are involved, such as data exploratory analysis and data pre-processing, which allows the creation and selection of the features compatible with the proposed method, boosted random forest. Additionally, with the newly created and selected feature, the model's performance will be advanced, and evaluations are made based on the confusion matrix and looking at the accuracy of the model.

3.2 Research design

Research design generally describes the logical sequence of the research study. The study focuses on developing a model that will predict whether a transaction made using a credit card is a fraud or not, based on the time and amount of the transaction. The predictions made by the model be beneficial to both companies like banks and its client since they will be able to detect and stop fraud from being occurring. Hence both customer and company will not be harmed.

There are many predictive models in machine learning that can be used to make such predictions. Machine learning is categorised into supervised and unsupervised. In supervised machine learning, we will train the machine by learning from

labelled data to make predictions, and this type of learning builds predictive models. Unsupervised machine learning makes predictions from unlabelled data, and they learn from the structure of unlabelled data to make predictions. The research will use supervised learning since the prediction of credit card frauds will classify transactions as either fraud or not a fraud.

There are two supervised learning techniques to solve linear or classification problems: linear regression and classification techniques; linear regression is a technique that is usually used in anticipating, modelling, and identifying quantitative data relations. In contrast, using data processing and pattern recognition, classification predicts a qualitative response. There are several classification methods, but some of the commonly used techniques are logistic regression, k-nearest neighbours (KNN), decision tree, random forest, artificial neural networks (ANN), and support vector machine (SVM). Random forest is selected for this study, wherein it will be combined with the Adaptive boosting (AdaBoost) algorithm to solve the classification problem. AdaBoost is introduced for better performance of a random forest model.

3.3 Data

3.3.1 Dataset information

The dataset was chosen from Kaggle; it is a simulated credit card transaction dataset that includes both valid and fraudulent transactions. It protects the credit cards of 1000 clients that transact with a pool of 800 merchants. Brandon Harris' Sparkov Data Generation GitHub tool was used to build the dataset. There are 1296675 observations and 23 variables in the dataset. The simulation was run from January 1st, 2019, to December 31st, 2020.

3.3.2 Exploratory data analysis

Exploratory data analysis will help analyse the data to determine their characteristics, such as the variable with the most robust prediction, such that we will be

able to identify it. The process involves cleaning data and removing irrelevant information from the data. The irrelevant information from our dataset could be any features that are unique identifiers such as name, last name, account numbers and any other information present on the dataset after the fraudulent activities have been executed, such as a year, merchant longitude and a zip code. We can also view and understand the relationship between our features, which also helps in understanding features that are important in our study.

3.4 Methods

3.4.1 System specification

The scientific packages used in this analysis are imported using Python 3. The model is implemented from Scikit learn, a python three(3) library. The python software runs on 16GB Lenovo ThinkPad with Windows 10 operating system.

3.4.2 Data pre-processing

Data preparation (also known as "data preprocessing") transforms raw data so that data scientists and analysts can use machine learning algorithms to discover information or make predictions. It is a key to solving a problem [57], And if done properly, it will increase the efficiency of a boosted random forest model [82]. However, it is crucial to understand the dataset structure before data preparation. Furthermore, the process is known as exploratory data analysis(EDA), EDA is a long-standing statistical tradition that offers conceptual and analytical tools for finding patterns in order to promote hypothesis formation and refinement [10]. Exploratory data analysis will help analyse the data to determine their characteristics, such as the variable with the most robust prediction, such that we will be able to identify it.

Furthermore, once we understand the dataset structure, we will undertake data cleaning, which is the process of fixing or eliminating incorrect, damaged, incorrectly formatted, duplicate, or incomplete data from a dataset. When integrating

many data sources, there are numerous opportunities for data to be duplicated or mislabeled. If the data is incorrect, the results and algorithms are untrustworthy, even if they appear correct. Because the techniques vary from dataset to dataset, it is impossible to prescribe the exact steps in the data cleaning method. Data cleansing is critical since it ensures that we have the essential data for predicting credit card fraud, increasing a model's accuracy.

Whenever the data is ready, we will implement the data preprocessing technique. Preprocessing is a technique that converts the raw data into an intelligible format. Raw data, known as real-world data, is often incomplete, and it cannot be transmitted through a model, which will make some errors, which is why data must be preprocessed before sending through a model. There are several steps involved in data preprocessing: importing libraries, reading the data, checking for missing values, checking for categorical data, scaling the data, and splitting the data. This is a stage where the feature creation process is carried out.

Feature creation is a way of creating new features from one or more existing features, which might be used in statistical analysis. This process will add new features accessible during a model, hopefully generating an accurate model. Every model requires features to have specific characters to perform well. The reason for performing feature creation is that it will help prepare a proper input dataset compatible with the random forest algorithm and improve the random forest model's performance, which will improve the performance of a model, reducing the chances of model over-fitting and running time.

For feature creation, in this study, we will use normalisation to create new features that will be scaled between a specific range using min-max normalisation. In min-max normalisation, a linear transformation is done on the original dataset, fetching the minimum and maximum values from the dataset and replacing each value according to the following formula.

$$X' = \frac{X - Min(A)}{Max(A) - Min(A)}$$

Where,
A is the attribute data
Min (A) is the minimum absolute value of A
Max (A) is the maximum absolute value of A
X' is the new value of each entry in the dataset
X is the original (old) value of each entry in the dataset.

The min-max normalisation techniques will range from 0 and 1 since our dataset contained different ranges of values. For example, gender has a range from 0-1, while the amount range from 0-1000000. In this case, limit balance is about 10000 times larger than age; with a dataset with different feature ranges, the model will not learn because different features have different ranges. To avoid this kind of difference between features, we will apply normalisation on the whole dataset such that the model produces efficient and accurate results.

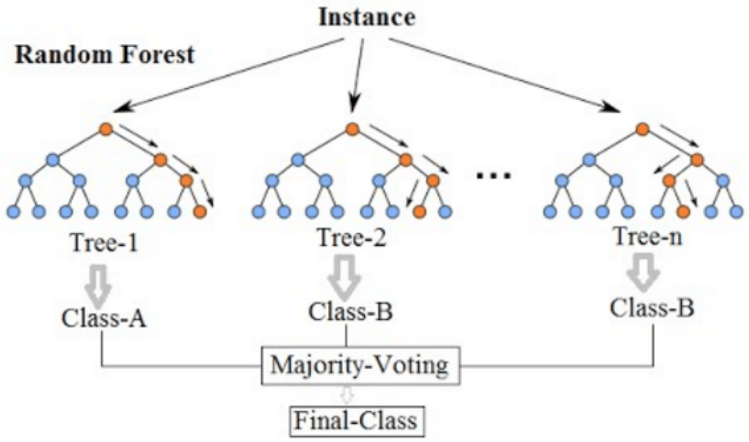


FIGURE 3.1: Random Forest

Random forest is a supervised algorithm of classification trees; it works for both regression and classification problems. Random forests create decision trees on randomly selected data samples, get predictions from each tree, and select the best solution through voting as indicated in Figure 3.1 [46]. When trees operate together, they protect each other from individual errors in this model. While some trees might be wrong, others will be right, so as a group, they can make the right decision, which will increase the chances of an accurate prediction; this is what makes the random forest model advantageous over other predictive models.

During the training, trees in random forests learn from random samples of data points. The random forest can memorise the training data wherein samples are drawn by a replacement, called bootstrapping. During testing time, predictions of each tree are generated by averaging predictions of each tree. Training each tree on different bootstrapped subsets of data and average the predictions is known as bootstrap aggregation. Bootstrap aggregation reduces the variance between learned and trained data since the random forest has a high variation between learned data and trained data.

This study will focus on the classification problem. Additionally, a random forest algorithm looks for the best suitable features among a set of features rather than the best possible features when splitting a node from multiple trees. Random forest model will be implemented from the scikit-learn library to perform feature importance for the feature selection process

The most critical thing we look at while building a model is the efficiency and accuracy, which are more important things that should be optimised. So we can improve the model's prediction accuracy by increasing the number of trees before making any computation on averaging predictions. Having more trees is disadvantageous since it decreases the speed of a model. Besides the increasing number of trees, we can also maximize the number of leaves that a model requires to split the node.

Other parameters can also help improve the performance of a random forest model. However, other random forest parameters we can use to improve the speed of the model are "N-JOBS," "RANDOM-STATE," and "OOB-SCORE."

(i) Feature selection
Feature selection helps solve over-fitting and the curse of dimensionality in a data set by reducing the number of features in a model and optimising the model's performance. Reducing the number of features, the output model will become more efficient and more comfortable to interpret and make; hence the model will make accurate feature predictions. We will use the feature importance rank since the random forest is selected for this study and is a tree-based model. Mean decrease impurity and mean decrease accuracy are used in feature importance random forest.

(ii) Random forest feature importance
We will use the random forest feature importance for feature selection, which will allow us to find the feature importance of each variable in the dataset. Feature importance will give us the score of each variable/feature in the dataset. The features with a high score are the once, which will be more important than the target variable's feature.

AdaBoost is a machine learning meta-algorithm that can enhance the efficiency of many other machine learning algorithms. The other machine learning algorithms ('weak learners') are combined into a weighted sum that represents the boosted classifier's final output [16]. AdaBoost is adaptive because subsequent weak learners are tweaked in favour of those instances misclassified by previous classifiers. In specific problems, it may be less susceptible to over-fitting than other machine learning algorithms. Individual learners can be poor, but as long as each output outperforms random guessing, the final model can be seen to converge to a strong learner. Adaptive Boost algorithms can be used for both classification and regression problems.
So, due to the extreme random training, the random forest is resistant to noise and highly generality. It does, however, necessitate a large number of decision trees, as using fewer decision trees lowers the efficiency of a model. As a result, it loses its generality when implemented on small-scale hardware. As a result, boosting has been added to the random forest.

Moreover, the feature importance technique is used for feature selection, which will allow us to find the feature importance of each variable in the dataset. Feature importance will give the score of each variable/feature in the dataset. The features

with a high score are the ones that will be important for our model towards credit card fraud detection.
The performance of our boosted random forest will be evaluated using model accuracy and confusion matrix. The confusion matrix is the table that describes the performance of a model on a set of test data for which actual values are known. The confusion matrix is primarily used in studies that solve the classification problem on unbalanced data.

3.5 Evaluation matrices

Evaluation matrices selected for this study are:

- Confusion Matrix
- Accuracy, Recall, Precision and F1-core
- ROC CURVE and AUC

3.5.1 Confusion matrix

This is the evaluation matrice that indicates which indicates the classes that were classified correctly, which classes are misclassified and what type of errors are made. Below is the metrics of confusion matrix:

TABLE 3.1: Confusion matrix

	True Positive	True Negative
Predicted Positive	TP	FP
Predicted Negative **1**	FN	TN

- True Positive (TP): The proportion of predictions in which the classifier correctly predicts the positive class as positive.

- True Negative (TN): The proportion of predictions in which the predictor adequately predicts that the negative class is negative.
- False Negative (FN): The proportion of predictions in which the predictor predicts a negative class as a positive class.
- False Positive (FP): The proportion of predictions in which the predictor wrongly anticipates a positive class as a negative class.

Some of the performance measures are used from the confusion matrix are:

- **accuracy**: It offers you the model's overall accuracy, which is the percentage of total samples correctly identified by the classifier.
- **Recall**: It indicates how many positive samples were predicted correctly as positive by the classifier.
- **precision**: It tells you what percentage of forecasts in the positive class was correct.
- **F1-score**: It incorporates precision and recalls into a single metric.
- **ROC CURVE and AUC**: This is of the evaluation matrices that represent the actual performance of the model considering all possible probability cut-offs.

3.6 Summary

Since this study is detecting whether a transaction made is fraudulent or not, this is a type of classification problem which classifies transactions based on two answers presented in the form of binaries or not. The research technique used will be a boosted random forest. It can make the detection of fraudulent behaviour reliable because of the boosting algorithm added to the random forest algorithm. The algorithm also uses an ensemble of trees in making decisions and can handle missing values, making it a model that promises good results. However, since it is time-consuming to make predictions, we will use high-importance features discovered in the feature selection process, reducing waiting time.

Chapter 4

Results analysis

4.1 Introduction

The study focuses on analysing the performance of boosted random forest on credit card fraud detection. The analysis is based on the financial data obtained from an online source called Kaggle. The results in this section include the dataset information and the visualisations which are then used to understand customer's spending behaviour, such that we can come up with insightful information that helps in understanding the cause of these fraudulent transactions and also what is it that we can look at to say the transaction is suspicious or fraudulent. Hence out of those insights, we can develop the model to detect fraudulent cases efficiently and effectively. Understanding customer behaviour improves the feature creation and feature selection process, for the purpose of improving the model's performance.

4.2 Exploratory data analysis

The dataset includes the variable "Unnamed" which in our study is not relevant, since this is the variable containing the automatically generated records. Additionally, other variables, "first" and "last", which indicate the first and the last name of a credit card holder, were also removed from the dataset because all these variables had no importance towards detecting credit card fraud. Moreover, apart from the three variables (Unnamed, first, last), there is no other containing any redundant information, poorly formatted or having no impact on the study. Therefore for analysis testing of the model, only 18 variables are left, with the last variable on the

dataset being the target variable.

Looking at the dataset information, we discovered the number of columns, number of records, number of entries, data-type and memory used by the dataset.

TABLE 4.1: Confusion matrix of random forest model on testing dataset

```
==================================================
Information About Dataset

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1604294 entries, 0 to 555718
Data columns (total 20 columns):
amt                     1604294 non-null float64
category                1604294 non-null object
cc_num                  1604294 non-null float64
city                    1604294 non-null object
city_pop                1604294 non-null int64
dob                     1604294 non-null object
gender                  1604294 non-null object
is_fraud                1604294 non-null int64
job                     1604294 non-null object
lat                     1604294 non-null float64
long                    1604294 non-null float64
merch_lat               1604294 non-null float64
merch_long              1604294 non-null float64
merchant                1604294 non-null object
state                   1604294 non-null object
street                  1604294 non-null object
trans_date_trans_time   1604294 non-null object
trans_num               1604294 non-null object
unix_time               1604294 non-null int64
zip                     1604294 non-null int64
dtypes: float64(6), int64(4), object(10)
memory usage: 257.0+ MB
None

==================================================
```

From Table 4.1 it indicates that there are 18 variables and 1604294 records, data types are: float 64 (6), int 64 (4), object (12) and memory used is 281.5 MB. Moreover, as part of data processing is also essential to check if there are any missing values; from Table 4.1 results shows that the dataset does not have any missing values.

Our study is supervised learning wherein the dataset is labelled, with the target variable indicating whether the transaction made by a customer was legit or not. Moreover, it is represented by binary numbers 0 and 1, where 0 represent legit (non-fraudulent transaction), and 1 represent non-legit(fraudulent transaction).

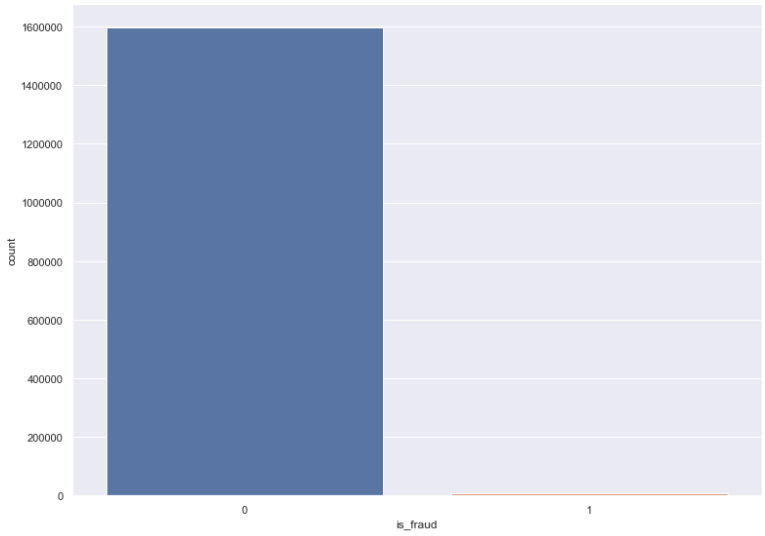


FIGURE 4.1: Non-fraudulent transactions (0) vs fraudulent transactions (1)

From Figure 4.1, the dataset contains a lot of legit transactions, wherein it has 1596143 non-fraudulent transactions and 8151 fraudulent. It is also supported by the diagram below, which shows the visualisation between 0 and 1's, wherein it can be seen that the dataset is highly imbalanced.

Since we are dealing with credit card fraud, it is vital to understand the customer's behaviour to get answers on how they spend their money.

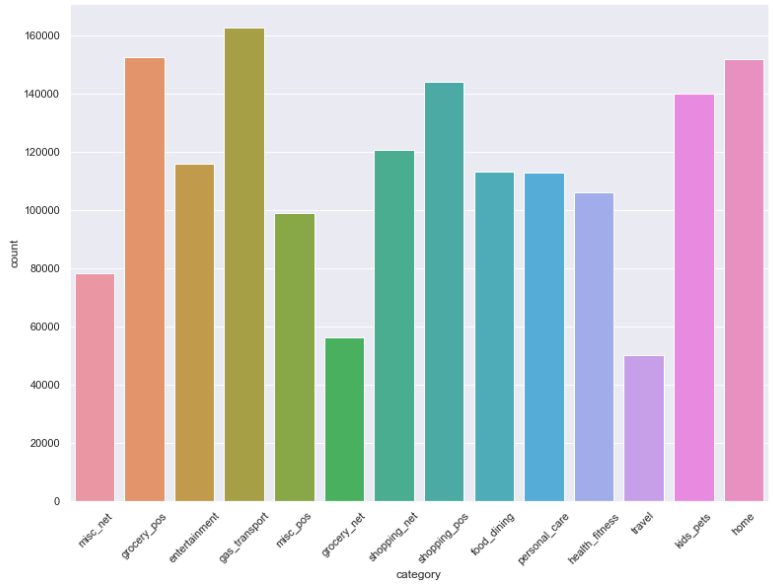


FIGURE 4.2: Merchants where customers use their credit cards

The Figure 4.2, most customers spend their money most on Gas and transport since there was the highest number of transaction, and spent less on travel since it was the lowest.

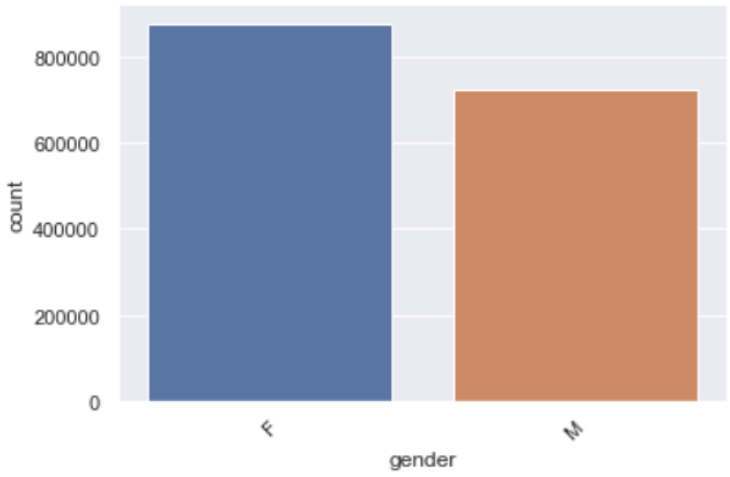


FIGURE 4.3: Female vs Male customers

Looking at the gender in which was making a lot of transactions using the credit card, Figure 4.3 shows that females make more transactions than males. And this also means that females are more likely to be victims of fraud since they make many transactions with a credit card.

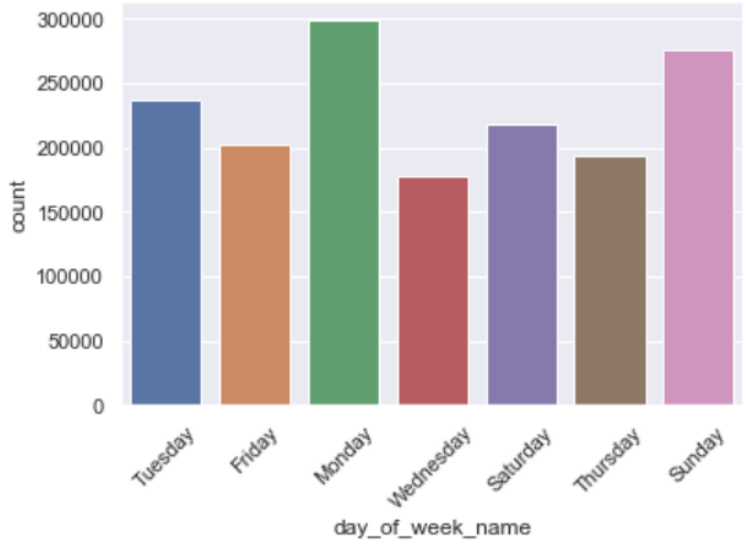


FIGURE 4.4: Days of the week.

A lot of transactions made by the customers were made during the night. And most of these transactions were made on Mondays and Sundays.

4.3 Feature creation

The dataset has values with different ranges. The dataset having different ranges tends to have many outliers, resulting in the model over-fitting (the model will not be able to learn). Hence, giving out the results is not accurate.

So, in our study, we create new features by scaling the whole dataset between the range on 0 and 1. The technique used to normalise the dataset is min-max normalisation, normalised using the default range.

4.4 Model results and performance

To evaluate the model's performance, we split the dataset into an 80:20 split ratio, where 80% of the dataset is for training, and 20% is for testing a model. Therefore, the training set contained 1283435 observations, and the testing set contained 320859 observations. The data in our study is not split randomly. We set the parameter random-state=0. The splitting of the dataset is done on the normalised data. The data in our study is not split randomly. We set the parameter random-state=0. The splitting of the dataset is done on the normalised data. In this study, we are dealing with a classification model. We will evaluate the model's performance using a confusion matrix; a confusion matrix is mostly used in a classification problem. Additionally, the confusion matrix enables us to find other performance measures: recall and precision of a random forest. A recall is the proportion of actual positive cases that are correctly identified, and precision is the proportion of positive cases that were correctly identified.

4.4.1 Random forest

In the random forest, we need as many trees as possible for our model to be better and produce accurate results. The random forest model is implemented using the Scikit learn python library, consisting of different built-in functions. Random forest models have different parameters that impact the performance of the model. Below is the python code showing different random forest parameters,

TABLE 4.2: Python code for random forest parameters

```
In [31]:
from sklearn.ensemble import RandomForestClassifier
RF=RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                          max_depth=10, max_features='auto', max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=3, min_samples_split=5,
                          min_weight_fraction_leaf=0.0, n_estimators=2000,
                          n_jobs=None, oob_score=False,
                          verbose=0, warm_start=False)
RF.fit(x_train,y_train)

Out[31]:
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=10, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=3, min_samples_split=5,
                       min_weight_fraction_leaf=0.0, n_estimators=2000,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

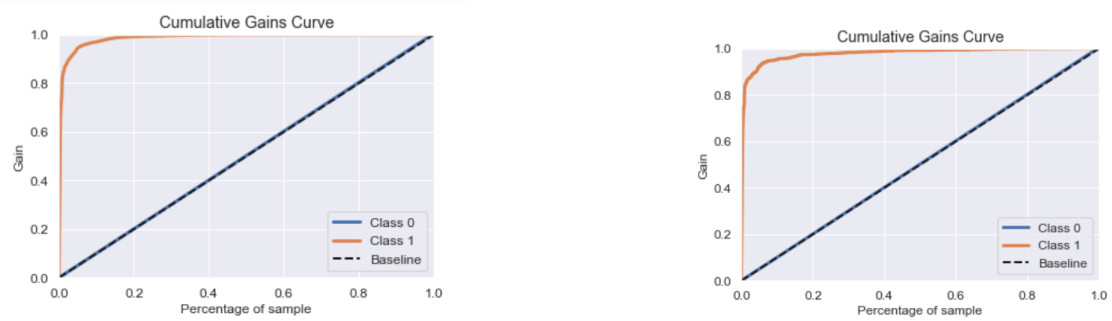
We changed some parameters to ensure better model performance, and those parameters that were changed are max-depth, min-sample split, min-sample leaf, and n-estimators. These parameters contributed to solving model over-fitting like producing 100% in training and 82% in testing accuracy and enabling the model to produce accurate results. The remaining parameters were set by default.

When training a random forest model, we must ensure that we have as many trees as possible. So in our study, we used "n-estimators=2000", which indicate the number of trees on Table 4.2, which we use in our training and testing our model. According to the confusion matrix in Table 4.3, there are 639 true positive and 319213 true negative values.

TABLE 4.3: Confusion matrix of random forest model on testing dataset

	True Positive	True Negative
Predicted Positive	319171	74
Predicted Negative	765	849

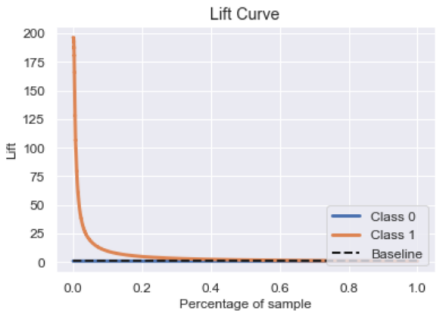
The random forest performed well in predicting the majority of the class (legit transaction represented by the 0) since out of 319245 legit transactions, 319171 transactions were predicted correctly. Looking at the minority class, which contains fraudulent transactions, random forest performed poorly; out of 1614 fraudulent transactions, only 849 were predicted correctly.



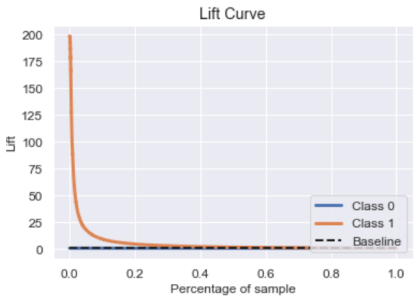
(A) Training set (B) Testing set

FIGURE 4.5: Gain charts for random forest model

Gain at a given decile level is the ratio of a cumulative number of targets up to that decile to the total number of targets in the entire data set, from both of our charts Figure 4.5a and 4.5b, for training and testing, the percentage of targets covered at a given decile level. For example, the top 20% of the dataset would contain approximately 100% of the fraudulent transactions, which means by just approving the 20% of the transaction, even the fraudulent transaction would be approved.



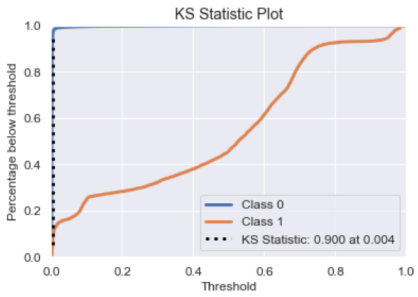
(A) Training set



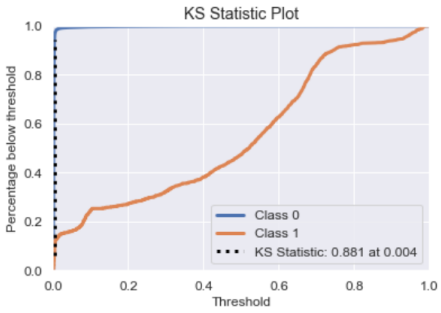
(B) Testing set

FIGURE 4.6: Lift charts for random forest model

Lift chart from Figure 4.6a and 4.6b, measures how much one can expect to do with a predictive model compared without a model. It is the ratio of the gain percentage to the random expectation percentage at a given decile level. From the testing set, the Lift of 25 for two deciles means that when selecting approximately 10% of the records based on the model, one can expect 25 times the total number of targets (events) found by randomly selecting approximately 10% without a model.



(A) Training set



(B) Testing set

FIGURE 4.7: Kolmogorov-simimov statistics for random forest model

Kolmogorov-simimov statistics (KS statistics) this is a metric used by credit card issuers to know how many transactions they should target in order to be able to detect fraudulent transactions. Our KS-statistics from the training set is 90%, and in the testing set, our KS-statistics is 81%.

4.4.2 Adaptive boosting

In the Adaptive boosting, we use same parameter with same values as random forest e.g. (n-estimators=2000).Below is the python code showing different random forest parameters,

TABLE 4.4: Python code for adaptive boosting parameters

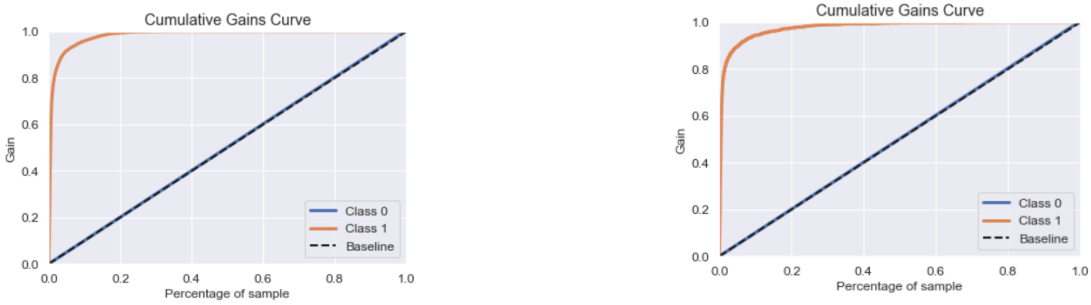
```
In [39]:
1 ab= AdaBoostClassifier(n_estimators=2000, learning_rate=1)
2 model=ab.fit(x_train,y_train)
3
4
```

According to the confusion matrix in Table 4.5, there are 472 true positive and 318966 true negative values.

TABLE 4.5: Confusion Matrix of adaptive boosting Model

	True Positive	True Negative
Predicted Positive	318969	276
Predicted Negative	1195	419

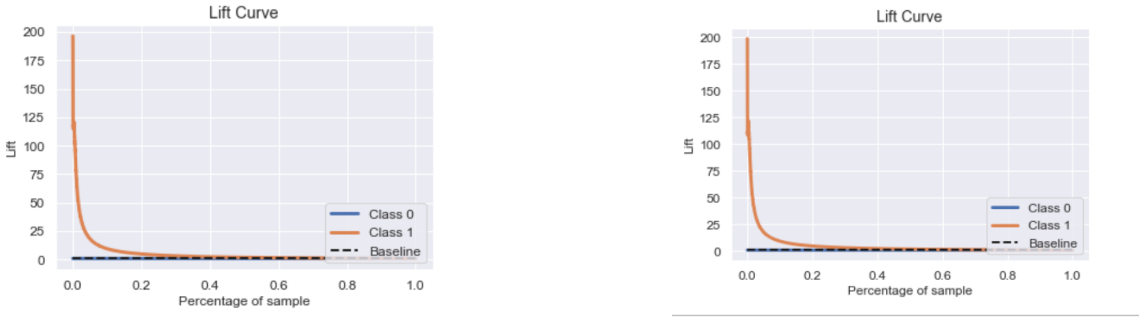
The adaptive boosting performed well in predicting the majority of the class (which is legit transaction represented by the 0) and the minority of the class compared to the random forest model since out of 319245 legit transactions, 318969 transactions were predicted correctly. Looking at the minority class, which contains fraudulent transactions, out of 1614 fraudulent transactions, only 419 were predicted correctly.



(A) Training set (B) Testing set

FIGURE 4.8: Gain chartsfor adaptive boosting model

Gain at a given decile level is the ratio of a cumulative number of targets up to that decile to the total number of targets in the entire data set, from both of our charts in Figure 4.5a and 4.5b, for training and testing, the percentage of targets covered at a given decile level. For example, the top 20% of the dataset would contain approximately 100% of the fraudulent transactions, which means by just approving the 20% of the transaction, even the fraudulent transaction would be approved.

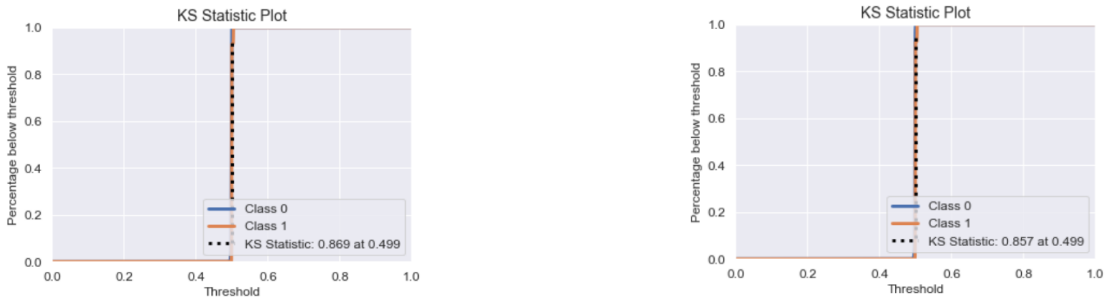


(A) Training set (B) Testing set

FIGURE 4.9: Lift charts for adaptive boosting model

Lift chart from Figure 4.9a and 4.9b, measures how much one can expect to do with a predictive model compared without a model. It is the ratio of the gain percentage to the random expectation percentage at a given decile level. From the testing set, the Lift of 100 for two deciles means that when selecting approximately 0% of the records based on the model, one can expect 100 times the total number of targets

(events) found by randomly selecting approximately 0% without a model.



(A) Training set (B) Testing set

FIGURE 4.10: Kolmogorov-simimov statistics for adaptive boosting model

Kolmogorov-simimov statistics (KS statistics) this is a metric used by credit card issuers to know how many transactions they should target in order to be able to detect fraudulent transactions. From the training set, our KS-statistics is 87%, and in the training set, our KS-statistics is 86%.

Therefore looking at the overall performance of these two models, random forest performed better than adaptive boosting. However, looking at the accuracy of each of these models, we can tell that all these models were overfitting, with random forest having an accuracy of 100%, whereas there is some prediction done incorrectly. Moreover, adaptive boosting with the accuracy of 99%. Furthermore, neither random forest nor adaptive boosting could correctly predict 100% or 99% of the dataset as per the confusion matrix in Table 4.3 and 4.5.

4.4.3 Boosted random forest

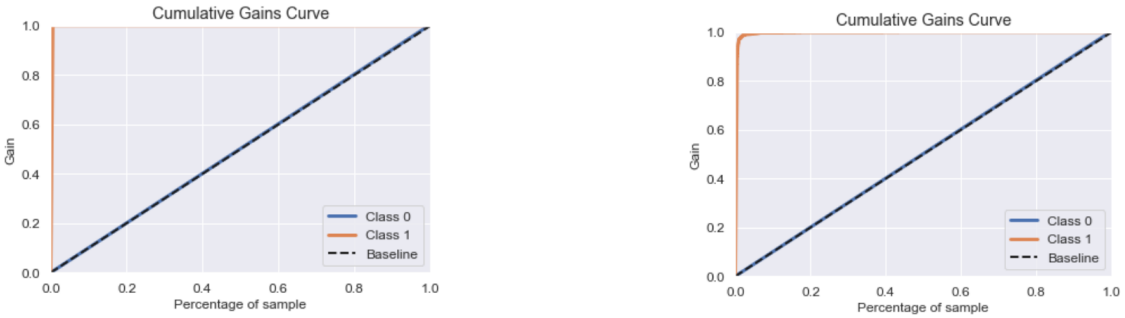
Random forest and Adaptive boosting in this study are combined to develop the best model, with the random being the base model and the adaptive boosting model, for boosting the performance of the random forest, ensuring that there is no overfitting. The boosting model is leveraged in the boosted random forest when a low variance and high bias are observed to ensure an optimal final model.

TABLE 4.6: Confusion matrix of boosted random forest model

	True Positive	True Negative
Predicted Positive	319190	55
Predicted Negative	295	1319

According to the confusion matrix in Table 4.6, there are 1319 true positive and 319190 true negative values.

Boosted random forest model performed well in predicting minority and majority classes. The boosted model in table 3, out of 319245 legit transactions on the dataset, 319190 transactions were predicted correctly. Looking at the minority class, which contains fraudulent transactions, out of 1614 fraudulent transactions, only 1319 were predicted correctly. Furthermore, the model's accuracy is 99.9%, which means that 99.9% of the overall dataset was predicted correctly, which is also scientifically proven by the results in Table 4.6.



(A) Training set (B) Testing set

FIGURE 4.11: Gain charts for boosted random forest model

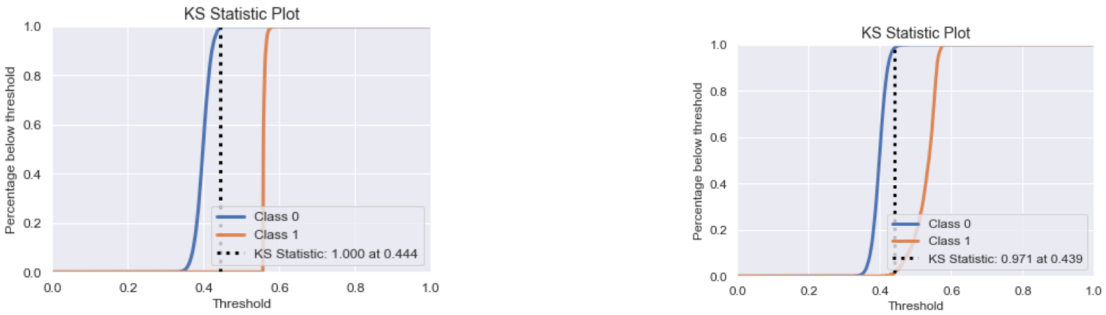
Gain at a given decile level is the ratio of a cumulative number of targets up to that decile to the total number of targets in the entire data set, from both of our charts Figure 4.5a and 4.5b, for training and testing, the percentage of targets covered at a given decile level. For example, the top 20% of the dataset would contain approximately 100% of the fraudulent transactions, which means by just approving the 20% of the transaction, even the fraudulent transaction would be approved.

(A) Training set (B) Testing set

FIGURE 4.12: Lift charts for boosted random forest model

Lift chart from Figure 4.12a and 4.12b, measures how much one can expect to do with a predictive model compared without a model. It is the ratio of the gain percentage to the random expectation percentage at a given decile level. From the testing set, the Lift of 100 for two deciles means that when selecting approximately 0% of the records based on the model, one can expect 100 times the total number

of targets (events) found by randomly selecting approximately 0% without a model.



(A) Training set (B) Testing set

FIGURE 4.13: Kolmogorov-simimov statistics for boosted random forest model

Kolmogorov-simimov statistics (KS statistics) this is a metric used by credit card issuers to know how many transactions they should target in order to be able to detect fraudulent transactions. From the training set, our KS-statistics is 87%, and in the training set, our KS-statistics is 86%.

The boosted random forest model performed better than the individual models (random forest and adaptive boosting). It also solved the issues of overfitting and being biased against the minority class, proving that the boosted random forest predicted those transactions accurately. Therefore, the boosted random forest is the best model for detecting credit card fraud.

Feature Selection

We used the Scikit learn library to compute feature selection using random forest feature importance. Below are the rankings showing how important a feature is in predicting credit card defaulters.

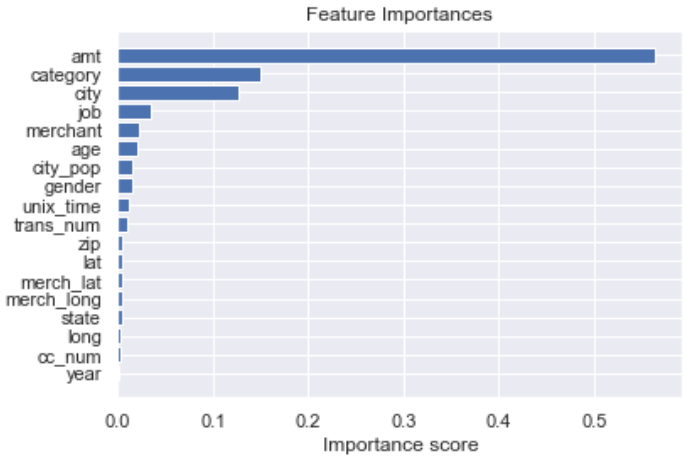


FIGURE 4.14: Random forest feature importance

Figure 4.14 helps us see the essential features that play a significant role in predicting credit card defaulters. From the graph, "amt = amount" is the strongest predictor since it has the highest importance score, which indicates that the amount of money the customer use daily can play a vital role in detecting whether the transaction made is fraudulent or not. Moreover, this is because if the customer makes a particular purchase of a certain amount between a specific range, and out of nowhere the customer decides to exceed that limit two times or more than what they usually spend, it raises concerns to the system and hence can be detected as a fraud.

Adding to the list of solid predictors of credit card fraud features is "category= indicating the merchant's customer likes spending their money) and "city = a place where the customer stays and usually spends their money". The top 3 features are predictors that the credit card system can rely on detecting fraudulent activities, which means if the transaction made the amount exceed what the customer usually spend. Moreover, made from different merchants where the customer usually spends their money and also from a different country, it indicates a high probability

that the transaction is fraudulent, hence rejecting the transaction.

In our study, we only removed the last six features from Figure 4.14: which are additional information for transaction and personal features (" long"," Unix-time", "Mech-lat", "Mech-long", "lat", "occ-num", "zip" and "Year") and these feature are only available on the data after the fraud has been committed. In total, for our model training and testing, we used ten features.

TABLE 4.7: Classification report for all models

	Precision	Recall	F1-Score	Roc AUC	Accuracy Rate
Random forest	96%	76%	83%	76%	100%
Adaptive boosting	80%	63%	68%	63%	99%
Boosted random forest	98%	91%	94%	91%	99.9%

Table 4.7 shows the model comparison based on Precision, Recall, F1-Score, ROC-AUC and the Accuracy Rate. The results show that the boosted random forest performed well in all instances; this proves that a boosted random forest model can effectively detect credit card fraud transactions.

4.5 Summary

The boosted random forest model, which combines the random forest and adaptive boosting, was developed; from the analysis of the results, the boosted random forest outperformed the individual model. Moreover, it is because of the valuable features created and selected according to how important they were towards credit card fraud detection, which resulted in the dataset having only features that were important towards detection of credit card fraud; hence the improved results from boosted random forest looking at both the accuracy and the confusion matrix compared to individual models.

Chapter 5

Summary, Conclusions and Recommendations

5.1 Summary

The study took into consideration the aim and the objectives of the study to answer the research questions. The models were developed and accessed in the detection of credit card frauds. The analysis and results obtained are discussed in chapter 4. This section gives a brief conclusion and recommendations for the feature work.

5.2 Conclusions

In this study, the three most essential objectives contributed to achieving the study's aim:to develop a model that will detect fraudulent credit card transactions accurately. The objectives were also used to prepare the dataset meant for training and testing the model.

Understanding the patterns of the data for this study, was crucial since we understood how customers spent their money, which also contributed towards looking at some exciting parts of the features that could be a source of interest to detect fraudulent transactions. We normalised our data and then had to look for features that were compatible with the model, such that we could ensure that the data fed to our model were relevant enough for our model to produce good results. We performed feature importance using the random forest to identify features with a high score in

credit card fraud detection.

We normalised our data and then had to look for features that were compatible with the model, such that we could ensure that the data fed to our model were relevant enough for our model to produce good results. We performed feature importance using the random forest to identify features with a high score in predicting credit card defaulters.

The results in Figure 4.14 feature "amt" which is the amount of money a customer spent using credit card" is the feature with a higher score than other most crucial feature. The feature "amt" shows customer spending behaviour, which can be used to determine whether the transaction made is fraudulent or not. Furthermore, it means that whenever a credit card transaction is made, the same customer's previous transaction must be checked to check if the amount used matches the customer spending behaviour by simply checking previous history transactions.

Besides the transactional amount, we can also look at the city where the transaction took place and the merchants where customers usually purchase things. Since we know that even the transaction is made far away or in a different city and from different merchants compared to what the credit card owner usually does, it can also assist in detecting fraudulent cases. Generally, we know that people usually make many transactions within the area they are staying. If made outside, it may be a fraudulent transaction, and then the payment will be stopped.

In detecting fraudulent transactions, we do not look into personal information like age, race, job and many other personal features. Moreover, anyone can be a victim of fraud regardless of their features. The most important thing to look at is the amount and place of transition and merchants where the transaction is taking place. Additionally, looking at Figure 4.14, it is seen that the last six features, which can be regarded as additional personal information and additional transitional information, are less critical in detecting fraudulent activities. Hence they were removed from the final dataset used to train and test the model.

With features created and selected; we developed a boosted random forest model

that could make detections efficiently and accurately compared to the individual models, which random forest and Adaptive boosting. The objectives were achieved, with the feature selected using random forest. We developed a model that detected credit card frauds at the accuracy of 99% with the supporting results in table 4.3, which is very good compared to the random forest and adaptive boosting with the accuracy of 100% and 99% with the supporting results in Table 4.3 and 4.5 respectively.

Our well-trained, boosted random forest model can improve it by finding suitable methods that deal with highly imbalanced datasets and analysing it on the different datasets. Since we behave with credit cards, people can differ in society, religion, race, tradition, culture, and environment. Our boosted random forest should still be accurate when detecting credit card fraud on a different dataset that changes with time and place.

5.3 Recommendations

The analysis of the study was done on the dataset, which is from an online source called Kaggle, and it contains information about people from outside Africa. So in future, we wish to perform some analysis on a different dataset since the behaviour of a customer with a credit card differs with different people of a different race, age, gender, and environment. Moreover, the boosted random forest will still detect credit card fraud accurately. We also encourage other researchers to analyse credit card fraud data to improve ways to manage fraudulent activities. Because credit card issuers give out credit cards since it is a significant investment. However, when credit card issuers do not have systems to monitor all the transactions fail, they suffer a huge loss. With many studies being done, institutions can identify a strategy to prevent fraudulent transactions.

Future research should focus on feature selection using an automated feature selection process since, in our study, we used a manual process to select features that were important in fraud detection. And also, feature studies should look at unsupervised learning toward the detection of credit card fraud, comparing the results

with a supervised model to see which models are best at fraud detection. Furthermore, this will help decide whether the supervised or unsupervised machine learning algorithm performs well in detecting credit card fraud, assisting in building future models for feature prevention of credit card fraud.

Bibliography

[1] Consumer Action. "Credit Card Fraud". In: (2009). URL: https://www.consumer-action.org/.

[2] Zafar U Ahmed et al. "Malaysian consumers' credit card usage behavior". In: *Asia Pacific Journal of Marketing and Logistics* 22.4 (2010), pp. 528–544.

[3] Nur Arifin Akbar et al. "Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm". In: *2020 International conference on informatics, multimedia, cyber and information system (ICIMCIS)*. IEEE. 2020, pp. 110–114.

[4] Emin Aleskerov, Bernd Freisleben, and Bharat Rao. "Cardwatch: A neural network based database mining system for credit card fraud detection". In: *Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr)*. IEEE. 1997, pp. 220–226.

[5] Doaa Almhaithawi, Assef Jafar, and Mohamad Aljnidi. "Example-dependent cost-sensitive credit cards fraud detection using SMOTE and Bayes minimum risk". In: *SN Applied Sciences* 2.9 (2020), pp. 1–12.

[6] K. Karuppasamy B. Mohankumar. "Credit Card Fraud Detection Using Random Forest Technique". In: *International Journal of Innovative Research in Science, Engineering and Technology* (2019). URL: www.ijirset.com.

[7] Bart Baesens, Sebastiaan Höppner, and Tim Verdonck. "Data engineering for fraud detection". In: *Decision Support Systems* 150 (2021), p. 113492.

[8] Alejandro Correa Bahnsen et al. "Feature engineering strategies for credit card fraud detection". In: *Expert Systems with Applications* 51 (2016), pp. 134–142.

[9] Alejandro Correa Bahnsen et al. "Feature engineering strategies for credit card fraud detection". In: *Expert Systems with Applications* 51 (2016), pp. 134–142.

[10] John T Behrens. "Principles and procedures of exploratory data analysis." In: *Psychological Methods* 2.2 (1997), p. 131.

[11] Mariana Belgiu and Lucian Drăguţ. "Random forest in remote sensing: A review of applications and future directions". In: *ISPRS journal of photogrammetry and remote sensing* 114 (2016), pp. 24–31.

[12] Tej Paul Bhatla, Vikram Prabhu, and Amit Dua. "Understanding credit card frauds". In: *Cards business review* 1.6 (2003), pp. 1–15.

[13] Siddhartha Bhattacharyya et al. "Data mining for credit card fraud: A comparative study". In: *Decision Support Systems* 50.3 (2011). On quantitative methods for detection of financial fraud, pp. 602–613. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2010.08.008. URL: https://www.sciencedirect.com/science/article/pii/S0167923610001326.

[14] Richard J Bolton, David J Hand, et al. "Unsupervised profiling methods for fraud detection". In: *Credit scoring and credit control VII* (2001), pp. 235–255.

[15] Rüdiger Brause, T Langsdorf, and Michael Hepp. "Neural data mining for credit card fraud detection". In: *Proceedings 11th International Conference on Tools with Artificial Intelligence*. IEEE. 1999, pp. 103–106.

[16] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[17] Bruno Buonaguidi et al. "Bayesian Quickest Detection of Credit Card Fraud". In: *Bayesian Analysis* 1.1 (2021), pp. 1–30.

[18] Joy Iong-Zong Chen and Kong-Long Lai. "Deep convolution neural network model for credit-card fraud detection and alert". In: *Journal of Artificial Intelligence and Capsule Networks* 3.2 (2021), pp. 101–112.

[19] Rong-Chang Chen et al. "Personalized approach based on SVM and ANN for detecting credit card fraud". In: *2005 International Conference on Neural Networks and Brain*. Vol. 2. IEEE. 2005, pp. 810–815.

[20] Linda Delamaire, Hussein Abdou, and John Pointon. "Credit card fraud and detection techniques: a review". In: *Banks and Bank systems* 4.2 (2009), pp. 57–68.

[21] Sahil Dhankhad, Emad Mohammed, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study". In: *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. 2018, pp. 122–125. DOI: 10.1109/IRI.2018.00025.

[22] Saurabh C Dubey, Ketan S Mundhe, and Aditya A Kadam. "Credit Card Fraud Detection using Artificial Neural Network and BackPropagation". In: *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE. 2020, pp. 268–273.

[23] Tom Fawcett and Foster Provost. "Adaptive fraud detection". In: *Data mining and knowledge discovery* 1.3 (1997), pp. 291–316.

[24] Javad Forough and Saeedeh Momtazi. "Ensemble of deep sequential models for credit card fraud detection". In: *Applied Soft Computing* 99 (2021), p. 106883.

[25] Sevdalina Georgieva, Maya Markova, and Velizar Pavlov. "Using neural network for credit card fraud detection". In: *AIP Conference Proceedings*. Vol. 2159. 1. AIP Publishing LLC. 2019, p. 030013.

[26] Sushmito Ghosh and Douglas L Reilly. "Credit card fraud detection with a neural-network". In: *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*. Vol. 3. IEEE. 1994, pp. 621–630.

[27] Sushmito Ghosh and Douglas L Reilly. "Credit card fraud detection with a neural-network". In: *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*. Vol. 3. IEEE. 1994, pp. 621–630.

[28] Md Al Mehedi Hasan et al. "Feature selection for intrusion detection using random forest". In: *Journal of information security* 7.3 (2016), pp. 129–140.

[29] Arvid OI Hoffmann and Cornelia Birnbrich. "The impact of fraud prevention on bank-customer relationships: An empirical investigation in retail banking". In: *International journal of bank marketing* (2012).

[30] Mia Huljanah et al. "Feature selection using random forest classifier for predicting prostate cancer". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 546. 5. IOP Publishing. 2019, p. 052031.

[31] K Indhumathi and K Sathesh Kumar. "Seasonal Infectious Disease Prediction based on Electronic Patient Health Records using Boosted Random Forest Algorithms". In: *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE. 2022, pp. 2025–2032.

[32] Celestine Iwendi et al. "COVID-19 patient health prediction using boosted random forest algorithm". In: *Frontiers in public health* 8 (2020), p. 357.

[33] Celestine Iwendi et al. "COVID-19 patient health prediction using boosted random forest algorithm". In: *Frontiers in public health* 8 (2020), p. 357.

[34] Yashvi Jain, ShripriyaDubey NamrataTiwari, and Sarika Jain. "A comparative analysis of various credit card fraud detection techniques". In: *International Journal of Recent Technology and Engineering* 7.5 (2019), pp. 402–407.

[35] M Kavitha and M Suriakala. "Real time credit card fraud detection on huge imbalanced data using meta-classifiers". In: *2017 international conference on inventive computing and informatics (ICICI)*. IEEE. 2017, pp. 881–887.

[36] Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. "Predicting the direction of stock market prices using random forest". In: *arXiv preprint arXiv:1605.00003* (2016).

[37] Kenji Kira and Larry A Rendell. "A practical approach to feature selection". In: *Machine learning proceedings 1992*. Elsevier, 1992, pp. 249–256.

[38] Byoung Chul Ko, Hyeong Hun Kim, and Jae Yeal Nam. "Classification of potential water bodies using Landsat 8 OLI and a combination of two boosted random forest classifiers". In: *Sensors* 15.6 (2015), pp. 13763–13777.

[39] Byoung Chul Ko, Hyeong Hun Kim, and Jae Yeal Nam. "Classification of potential water bodies using Landsat 8 OLI and a combination of two boosted random forest classifiers". In: *Sensors* 15.6 (2015), pp. 13763–13777.

[40] Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC, 2019.

[41] M. S. Kumar et al. "Credit Card Fraud Detection Using Random Forest Algorithm". In: *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*. 2019, pp. 149–153. DOI: 10.1109/ICCCT2.2019.8824930.

[42] M Suresh Kumar et al. "Credit card fraud detection using random forest algorithm". In: *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*. IEEE. 2019, pp. 149–153.

[43] Miron B Kursa and Witold R Rudnicki. "Feature selection with the Boruta package". In: *Journal of statistical software* 36 (2010), pp. 1–13.

[44] J Lee. "Credit card fraud will increase due to the Covid pandemic, experts warn". In: *CNBC* (2021).

[45] Chenglong Li et al. "Comparative study on credit card fraud detection based on different support vector machines". In: *Intelligent Data Analysis* 25.1 (2021), pp. 105–119.

[46] Gui Liyu. "Application of Machine Learning Algorithms in Predicting Credit Card Default Payment". In: *UCLA Electronic Theses and Dissertations* (2019). URL: https://escholarship.org/uc/item/9zg7157q.

[47] Wangchao Lou et al. "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes". In: *PloS one* 9.1 (2014), e86703.

[48] Yvan Lucas et al. "Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs". In: *Future Generation Computer Systems* 102 (2020), pp. 393–402.

[49] Sam Maes et al. "Credit card fraud detection using Bayesian and neural networks". In: *Proceedings of the 1st international naiso congress on neuro fuzzy technologies*. 2002, pp. 261–270.

[50] Yohei Mishina et al. "Boosted random forest". In: *IEICE TRANSACTIONS on Information and Systems* 98.9 (2015), pp. 1630–1636.

[51] Fatemeh Nargesian et al. "Learning Feature Engineering for Classification." In: *Ijcai*. Vol. 17. 2017, pp. 2529–2535.

[52] Vijayshree B Nipane et al. "Fraudulent detection in credit card system using SVM & decision tree". In: *International Journal of Scientific Development and Research (IDSDR)* 1.5 (2016), pp. 590–594.

[53] Suvasini Panigrahi et al. "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning". In: *Information Fusion* 10.4 (2009), pp. 354–363.

[54] Suvasini Panigrahi et al. "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning". In: *Information Fusion* 10.4 (2009), pp. 354–363.

[55] Raghavendra Patidar, Lokesh Sharma, et al. "Credit card fraud detection using neural network". In: *International Journal of Soft Computing and Engineering (IJSCE)* 1.32-38 (2011).

[56] Manuela Pulina. "Consumer behaviour in the credit card market: a banking case study". In: *International Journal of Consumer Studies* 35.1 (2011), pp. 86–94.

[57] Dorian Pyle. *Data preparation for data mining*. morgan kaufmann, 1999.

[58] Jon TS Quah and M Sriganesh. "Real-time credit card fraud detection using computational intelligence". In: *Expert systems with applications* 35.4 (2008), pp. 1721–1732.

[59] Jon TS Quah and M Sriganesh. "Real-time credit card fraud detection using computational intelligence". In: *Expert systems with applications* 35.4 (2008), pp. 1721–1732.

[60] JB Quraisha. "Stripped wallets, ripped hearts victims of financial fraud: An analysis beyond financial fraud". In: *International Journal of Social Sciences* 4.3 (2019), pp. 1101–1112.

[61] Arun Kumar Rai and Rajendra Kumar Dwivedi. "Fraud Detection in Credit Card Data Using Machine Learning Techniques". In: *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*. Springer. 2020, pp. 369–382.

[62] K RamaKalyani and D UmaDevi. "Fraud detection of credit card payment system by genetic algorithm". In: *International Journal of Scientific & Engineering Research* 3.7 (2012), pp. 1–6.

[63] Nilson Report. "Card Fraud Losses Dip to 28.58 Billion". In: *Nilson Report* (2022).

[64] Paulo Angelo Alves Resende and André Costa Drummond. "A survey of random forest based methods for intrusion detection systems". In: *ACM Computing Surveys (CSUR)* 51.3 (2018), pp. 1–36.

[65] Gabriel Rushin et al. "Horse race analysis in credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree". In: *2017 systems and information engineering design symposium (SIEDS)*. IEEE. 2017, pp. 117–121.

[66] Daniel Sánchez et al. "Association rules applied to credit card fraud detection". In: *Expert systems with applications* 36.2 (2009), pp. 3630–3640.

[67] Daniel Sánchez et al. "Association rules applied to credit card fraud detection". In: *Expert systems with applications* 36.2 (2009), pp. 3630–3640.

[68] Mukesh Saraswat and KV Arya. "Feature selection and classification of leukocytes using random forest". In: *Medical & biological engineering & computing* 52 (2014), pp. 1041–1052.

[69] Nishant Sharma. "Credit Card Fraud Detection Predictive Modeling". In: (2020).

[70] Aihua Shen, Rencheng Tong, and Yaochen Deng. "Application of classification models on credit card fraud detection". In: *2007 International conference on service systems and service management*. IEEE. 2007, pp. 1–4.

[71] Ajeet Singh and Anurag Jain. "Adaptive credit card fraud detection techniques based on feature selection method". In: *Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018*. Springer. 2019, pp. 167–178.

[72] Amit Singh, Ranjeet Kumar Ranjan, and Abhishek Tiwari. "Credit Card Fraud Detection under Extreme Imbalanced Data: A Comparative Study of Data-level Algorithms". In: *Journal of Experimental and Theoretical Artificial Intelligence* (2021), pp. 1–28.

[73] Priyanka Singh, Neha Katiyar, and Gaurav Verma. "Retail shoppability: The impact of store atmospherics & store layout on consumer buying patterns". In: *International journal of scientific & technology research* 3.8 (2014), pp. 15–23.

[74] Salvatore Stolfo et al. "Credit card fraud detection using meta-learning: Issues and initial results". In: *AAAI-97 Workshop on Fraud Detection and Risk Management*. 1997, pp. 83–90.

[75] Emma VA Sylvester et al. "Applications of random forest feature selection for fine-scale genetic population assignment". In: *Evolutionary applications* 11.2 (2018), pp. 153–165.

[76] K Tuyls, S Maes, and B Vanschoenwinkel. "Machine learning techniques for fraud detection". In: *VUB, Brussels, Belgium* (2000).

[77] Eesita Sen Vaishnave Jonnalagadda Priya Gupta. "Credit card fraud detection using Random Forest Algorithm". In: *International Journal of Advance Research, Ideas and Innovations in Technology* (2019). URL: www.IJARIIT.com.

[78] Christopher Whitrow et al. "Transaction aggregation as a strategy for credit card fraud detection". In: *Data mining and knowledge discovery* 18.1 (2009), pp. 30–55.

[79] Staff Writer. *These are the most common banking scams in South Africa*. URL: https://businesstech.co.za/news/banking/409869/these-are-the-most-common-banking-scams-in-south-africa/.

[80] Ren-Min Yang et al. "Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem". In: *Ecological Indicators* 60 (2016), pp. 870–878.

[81] Ong Shu Yee, Saravanan Sagadevan, and Nurul Hashimah Ahamed Hassain Malim. "Credit card fraud detection using machine learning as data mining technique". In: *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10.1-4 (2018), pp. 23–27.

[82] Shichao Zhang, Chengqi Zhang, and Qiang Yang. "Data preparation for data mining". In: *Applied artificial intelligence* 17.5-6 (2003), pp. 375–381.

[83] Xinwei Zhang et al. "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture". In: *Information Sciences* 557 (2021), pp. 302–316.

[84] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[85] Qifeng Zhou, Hao Zhou, and Tao Li. "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features". In: *Knowledge-based systems* 95 (2016), pp. 1–11.

[86] Lin Zhu et al. "A study on predicting loan default based on the random forest algorithm". In: *Procedia Computer Science* 162 (2019), pp. 503–513.