National College of Ireland

Feature Selection and Machine Learning Algorithms for an Improved Credit Card Fraud Detection System

MSc Research Project
Data Analytics

Boluwatife Joseph Omoworare
Student ID: 20185944

School of Computing
National College of Ireland

Supervisor: Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Boluwatife Joseph Omoworare

Student ID: 20185944

Programme: Data Analytics **Year:** 2021

Module: Master of science research project

Supervisor: Hicham Rifai

Submission Due Date: 16/12/2021

Project Title: Feature Selection and Machine Learning Algorithms for an Improved Credit Card Fraud Detection System

Word Count: 6318 **Page Count** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date: 16/12/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	

Date:	
Penalty Applied (if applicable):	

Feature Selection and Machine Learning Algorithms for an Improved Credit Card Fraud Detection System

Boluwatife Joseph Omoworare

20185944

Abstract

Our world today is characterised by the fast adoption of technology in everyday living. This is evident in the rise of e-commerce, online shopping, online gaming, e-learning, online payment systems, e-banking, etc. Therefore, it has become imperative that more research be conducted to investigate and develop more efficient and effective fraud detection systems. Several researches have leveraged the use of machine learning algorithms and complex data mining methodologies to produce efficient fraud detection systems. Hence, this paper implemented different machine learning fraud detection models with and without feature selection to ascertain if using feature selection improved the accuracy of the models. Therefore, correlation matrix was selected as the feature selection technique which was used against random forest and logistic regression. The dataset used was a synthetic dataset created by a researcher in the IBM TJ Watson Research Centre to address the lack of labelled datasets for credit card fraud detection. After implementation and evaluation, results showed that the Random Forest classifier with feature selection achieved the highest accuracy of 98%. Thus, showing that feature selection improved the accuracy of the models.

Keywords: *credit card fraud, feature selection, machine learning*

1. INTRODUCTION

With advancements in technology, traditional means of conducting commercial transactions are continuously being phased out and replaced with online-based transaction systems which are often carried out using credit cards. This has opened more avenues for cybercriminals who aim to use deceptive means to get end users to disclose their card details. They also use malicious activities to gain unauthorized access to the information being controlled by financial institutions. A paper by (Correa Bahnsen et al., 2016) revealed in 2018 credit card fraud was at about 45%, which increased to about 60% in 2016. Thus, credit card fraud can be described as the unauthorized usage of an individual's credit card details in carrying out financial transactions. These occurrences have caused great financial losses to financial institutions. (Pozzolo, 2015) reported that between 2007 and 2008, over $3 billion was lost to credit card fraud with over $5 billion in 2012.

Traditional means of detecting fraud such as expert rules have been recognized to be inadequate as criminals are also constantly coming up with new ways to go undetected. (Van Vlasselaer et al., 2015). Also, certain limitations such as fixed algorithms that are never modified have been identified with current machine learning techniques (Pozzolo *et al.*, 2014). Therefore, several researches have been conducted and more still ongoing to develop more effective fraud detection systems. These approaches leverage machine learning classifiers which have proven to be efficient in handling massive streams of data. This is characteristic of financial transaction data. These classifiers also have the capacity to scan transaction data at a fast rate to differentiate between fraudulent and non-fraudulent transactions. Additionally, it is important for fraud detection systems to be accurate and timely in detecting fraudulent transactions as sometimes fraud prevention measures fail in preventing such. Therefore, detection systems should be able to trigger alerts when a fraudulent transaction is carried out regardless of the prevention system/measures in place. This ensures such action is detected and acted upon in good time. As such more detrimental consequences are averted as most fraudulent transactions involve completely emptying funds in an account.

Subsequently, several machine learning techniques including supervised and unsupervised learning are applicable to detecting credit card fraud systems. The accuracy of these techniques has been identified to be greatly impacted by the feature selection technique employed. As such this paper applied a feature selection technique against two machine learning algorithms. Also, it highlighted that correlation-based feature selection against random forest produced more accurate results. Therefore, this research adopted the use of correlation matrix as the feature selection technique against random forest and logistic regression to produce a model that accurately detects credit card fraud. Consequently, it answered the Research Question **(RQ)**: **How accurate are machine learning algorithms using feature selection techniques for detecting credit card fraud?**
Hence the objectives of the paper are:

- To research current credit card detection techniques
- To build a model for detecting credit card fraud based on machine learning algorithms and feature selection technique
- To evaluate the accuracy of the models developed

Furthermore, its contributions include:

- Providing an overview of credit card fraud detection using a dataset of financial records
- Investigating current credit card detection measures issues and solutions
- Contributing to the research space by investigating the potential that lies in machine learning and feature selection for detecting credit card fraud.

The paper is organized accordingly. The next section presents related works on credit card fraud detections. It covers an overview of credit card fraud detection techniques, machine learning for credit card fraud detection, and feature selection techniques. This is followed by Section 3 which outlines the research methodology and section 4 which gives the design specification. A detailed explanation of the implementation is contained in

Section 5 while Section 6 gives an evaluation of the models. The research is concluded in Section 7 and also contains the limitations and proposed future works.

2. RELATED WORK

In (Lucas et al, 2019), the process of credit card fraud detection is defined as the process of revealing and preventing credit card fraud. It highlighted that it involves two steps which are the blocking time and the checking time. The blocking time is aimed at detecting fraudulent transactions while the checking time is aimed at detecting fraudulent cards. Therefore, blocking time should be shorter as they correspond to authorizing or not authorizing the financial transaction being conducted using a credit card. Consequently, it is important that fraud detection systems have high accuracy to be precise in determining which transactions are fraudulent. As such subsequent sub-sections will discuss how machine learning is being adopted for credit card fraud detection.

2.1 Overview of Credit Card Fraud Detection Techniques

Several researchers have developed credit card fraud detection models. (Suman, 2004) presented that these models are majorly based on supervised and unsupervised learning. Supervised learning techniques, for example, decision tree, random forest, neural network, and logistic regression classifiers are used to build models that could classify transactions into legitimate and fraudulent transactions, thus, detecting fraud. These models were trained using previously identified legitimate and fraudulent credit card transaction records (Raj and Portia, 2011). On the other hand, with unsupervised learning, an unlabeled dataset is used. These datasets represent normal account behavior which the model learns. Subsequently, the model then classifies any transaction that goes outside the set normal account behavior as fraudulent. (Carcillo et al., 2019) utilized unsupervised learning to develop a model that was able to detect credit card fraud. They reported that this method was complex and involved a long time for the model to learn every user's profile being controlled by a financial institution.

Furthermore, (Wen-Fang YU and Na Wang, 2009) used outlier mining and Distance sum algorithms as another technique for credit card fraud detection. In their experiment, they were able to accurately predict fraudulent transactions using credit card transaction dataset of a commercial bank. In (Bentley et al, 2000), a genetic algorithm was proposed for the detection of credit card fraud. This algorithm is based on genetic programming which is used to define logic rules that differentiate legitimate credit card transactions from fraudulent ones. (Altman et al, 2019) highlighted that although this proved to be successful in credit card fraud detections and reduced false alerts, it however, was still characterized by misclassifications. Another technique that received much attention was neural networks. An example is an online credit card fraud detection system developed in (Dorronsoro, 1997) which used neural networks to build a classifier that could detect fraud. This model was reported to be constrained by the need for data to be clustered by type of account. (Ezawa & Norton, 1996) also mentioned Bayesian networks as techniques suitable for the detection of credit card fraud. (Maes et al., 2002) suggested that although this technique had great potential, it was constrained by time especially when compared to neural networks

2.2 Common Challenges in Credit Card Fraud Detection

Several challenges exist in developing credit card fraud detection systems. These have been identified by some researchers and are thus highlighted below.

• **Unbalanced Data**: Data is a vital component in building any machine learning classifier model. Due to the nature of credit card transaction data, which usually is unequalled. This means that it generally contains extremely few amounts of fraudulent records across all card purchases. Therefore, classification is extremely complex and difficult in identifying fraudulent transactions, thus leading to inaccurate classifications (Samaneh et al., 2016).

Varied Relevance of Misclassification: There are two common measures for classification, these are false positive and false negative. The challenge describes how differently the impact of these various misclassification types is perceived by each financial institution. Errors in classification are generally grouped into Type I and Type II errors (James et al., 2013). Type I error refers to a false positive which indicates a regular transaction as fraud while a type II error refers to a false negative which indicates a fraudulent transaction as a normal transaction. These misclassifications are rated differently based on their effect on company performance. For example, Type I errors may not be rated as highly as type II errors as they are not so detrimental as this could be re-verified by the customer, unlike type II errors that can result in huge financial losses to the victim (James et al., 2013).

• **Failure to Adapt**: As new detection systems are developed, so also do fraudsters come up with new ways to evade them. These dynamic nature of fraudster techniques mean that classifier models are met with new forms of legitimate or fraudulent behaviors. Consequently, these detection models become ineffective in identifying new patterns of legitimate and fraudulent behavior. Regarding this, (Maes et al., 2002) emphasized the need for fraud detection models built on machine learning algorithms need to be constantly updated. This is to tackle the ever-changing nature and pattern associated with identifying legitimate and fraudulent transactions.

2.3 Machine Learning for Credit Card Fraud Detection

Machine learning algorithms have been around since the early 1960s (Jordan and Mitchell, 2015) and have been widely used in credit card fraud detection. It has been seen to provide more effective models that can learn and distinguish between fraudulent and legitimate credit card transactions. With machine learning, classifier models are trained using a dataset and an algorithm such as logistic regression, support vector machine, random forest, decision tree, and neural networks. The model learns from this input dataset and can classify or predict on an unknown dataset. Consequently, such machine learning algorithms have been used to develop credit card fraud detection models. For example, (Sahin and Duman, 2011) developed two machine learning models using Support Vector Machine (SVM) and decision tree for detecting credit card fraud using a real-world dataset. Their experiment evaluated the two models and concluded that the decision tree classifier outperformed the SVM

classifier in accurately predicting credit card fraud. In the same line, (Monedero et al., 2012) also developed a credit card fraud detection model based on the decision tree algorithm. The model was easy to implement and after evaluation, it was seen to have achieved high accuracy and was flexible in detecting features for each classification type. However, the research reported that using a decision tree required high computational resources and was highly time-consuming due to the number of branches. Another research that utilized machine learning for credit card fraud detection was seen in (Pun, 2011). Here, the researcher applied the use of the Bayesian network to classify credit card transactions into legitimate or fraudulent transactions. The dataset contained labelled credit card financial transactions which were used to train and test the model. After evaluation, the researcher identified that the Bayesian network achieved a high processing and detection rate. However, it involved extensive training which was time-consuming.

Furthermore, in a research conducted by (Xuan, et al, 2018), the random forest was used against a real-life dataset containing credit card transactions. The researchers concluded that the machine learning algorithm achieved good results when applied to a smaller set of data. However, the imbalanced nature of the dataset resulted in some problems. With the successes of these researches, this paper also utilized machine learning algorithms and a feature selection technique to develop classifier models that achieve high accuracy in detecting credit card fraud.

2.4 Feature Selection

Feature selection is an important aspect of machine learning as it greatly affects how the machine learning model performs. Here critical features from a dataset are selected to train the classifier model. This has been seen to improve the performance of the machine learning model. Other benefits of this step include reducing overfitting which means reducing the noise in a dataset. It also reduces training time.

As such, it is important to discover the relevant features in a dataset using feature selection. This will help to provide better outcomes of the classifier model. In developing credit card fraud detection models, feature selection techniques are used to find out the most relevant features that best distinguish between legitimate and fraudulent transactions. Several techniques exist for feature selection, some examples are correlation-based feature selection, recursive feature elimination, and univariate selection. In recursive feature elimination, each column is measured based on its importance in relation to the labelled column. Here, the columns with the least importance are eliminated. The process is repeated until only those columns which are important are identified. (Yan and Zhang, 2015) identified that this feature selection technique improved the effectiveness of a binary classifier when trained with synthetic gas sensor data. However, it was found to have some biases when used in the SVM classifier.

Additionally, (Emura, Matsui, and Chen, 2019) used univariate selection. Here, statistical tests are applied to identify the top features in the dataset. Each column is then assigned a score according to how they performed in identifying the labelled column. (Karegowda, Manjunath and Jayaram, 2010) applied the correlation-based feature selection technique in their experiment. Here, the correlation between each column and the labelled column is measured. A correlation matrix is produced which shows and ranks the relationship between all

features. With this, the features with correlation to the labelled column can be deduced as well as how high or low the correlation is. (Kumar, et al., 2019) also used confusion matrix and random forest in developing a credit card fraud classifier model. After evaluating the model, they reported that the performance of the model using the correlation-based feature selection achieved better results of about 90% accuracy. This paper also used this feature selection technique to identify features that are important in modelling. The next section highlights this paper's research methodology.

3. RESEARCH METHODOLOGY

In achieving the objectives of any research, it is important to follow a suitable methodology. As such, this research adopted the Knowledge Discovery in Databases (KDD) methodology. This methodology consists of iterative stages which are aimed at pattern and knowledge discovery from large datasets. (Fayad et al, 2020) highlighted that this process focuses on knowledge extraction from datasets and the efficient application of algorithms on large datasets. The diagram below shows the step-by-step process of the KDD methodology.

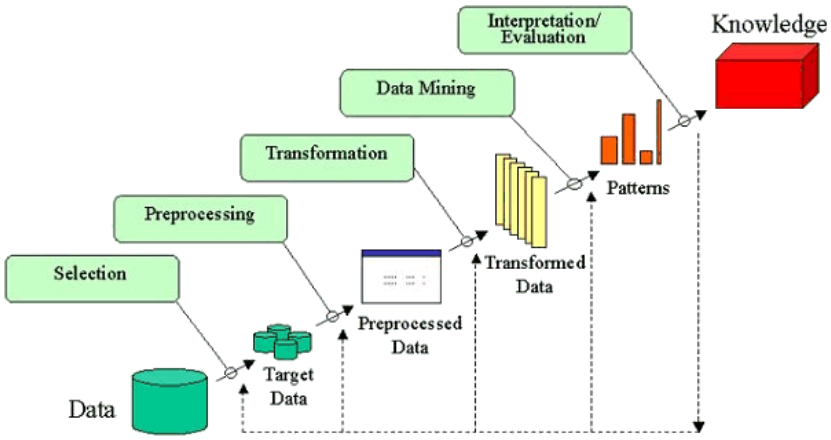


Figure 1: KDD METHODOLOGY

3.1 The Process Flow

To achieve the objectives of this research, certain steps were followed in alignment with the chosen research methodology. This project developed a classification model based on feature selection techniques and machine learning algorithms. The entire process flow starts with data collection, then data processing, feature selection, training, testing, and evaluation. The Figure below shows the process flow. Each step is discussed subsequently.

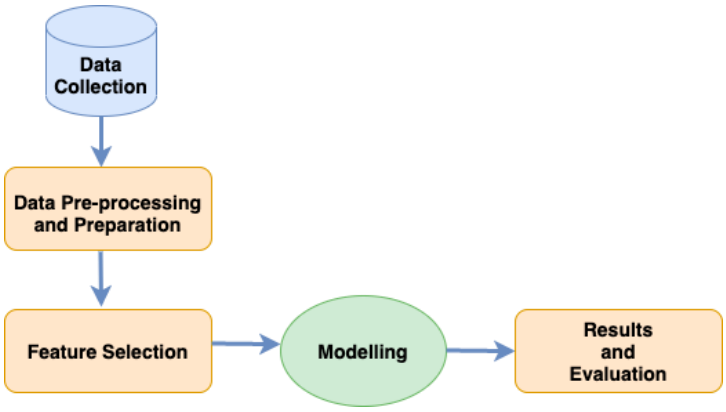


Figure 2: The Process Flow Diagram

3.1.1 Data Collection

One of the most vital steps in any machine learning project is the collection of data. The data collection step involves identifying and getting the dataset that will be used in training and testing the classifier model. In this project, a dataset was identified as suitable for achieving the aim of the research. The dataset used was a synthetic dataset containing over 24 million records and 15 columns. The columns include transaction date, amount, chip type, merchant location, and the labelled column. This dataset was created by a researcher in the IBM TJ Watson Research Centre (Altman, 2019) to address the lack of labelled datasets for credit card fraud detection. The dataset which was also created by simulating real-world financial transactions was downloaded from a publicly accessible repository (Altman, 2019). Furthermore, this research took into cognizance data privacy laws and ensured that the dataset was free for public use.

3.1.2 Data Pre-processing

Most public dataset available for research purposes usually contains noise such as missing data, null values, errors, etc. Hence, it is important that the dataset is pre-processed and prepared to remove any noisy data, missing values, and errors. This will ensure that a cleaner version of the dataset is used for modelling (García et al., 2015). The data pre-processing and preparation stage generally involves carrying out data normalization, data transformation, data reduction, data splitting, etc. These steps were considered in this paper after which feature selection was done.

3.1.3 Feature Selection

After all noisy data were eliminated from the dataset, the next important step was the selection of features relevant to training the classifier model. Feature selection is a process that ensures only those features that best help in classifying a transaction is selected. (Hall, 1999) stated that this phase identifies the features which are most relevant in modelling. While there are several techniques used in feature selection, this paper used Correlation-based feature selection. (Duangsoithong and Windeatt, 2010) reported in their paper that the

correlation-based feature selection was seen to have eliminated features that were irrelevant, thus resulting in a model that achieved higher accuracy when compared to another feature selection technique. Consequently, this paper used the correlation matrix as the correlation-based feature selection method which resulted in the elimination of irrelevant features in the dataset i.e., features that had low or no correlation with the labelled column.

3.1.4 Data Modelling

After the feature selection process, the machine learning algorithms were applied for training and testing. This defines the modelling phase. Here, the machine learning algorithm identifies patterns from the dataset and can predict them. As such, this paper used random forest and logistic regression as the two machine learning algorithms for training and testing the credit card fraud detection model. The dataset was split into the ratio 80:20 for training and testing respectively. Therefore, 4 models were developed.

3.1.5 Evaluation

Evaluating all developed models is important to ascertain their performance. As such, all four credit card fraud detection models developed were evaluated to determine how accurate they are and how they perform in detecting credit card fraud. Accuracy was taken as the primary criteria as it shows how correctly the developed models can classify credit card transactions into fraudulent and legitimate. Additionally, the recall, precision, f1-score and

Area Under Curve (AUC) was also recorded.

3.2 Proposed Algorithms

This research implemented credit card fraud detection models based on two machine learning algorithms. These were chosen as they have been recorded to be suitable for binary classification as well as accurate in classifying complex data. These algorithms are discussed below.

3.2.1 Random Forest

One of the most used machine learning algorithms in any classification problem is Random Forest. It has been applied in areas such as malware detection, anomaly detection as well as credit card fraud detection.

According to (Louppe, 2014), the random forest has been used in different areas with accurate prediction results. This is because it is more stable and efficient due to its creation of several trees (Louppe, 2014).

3.2.2 Logistic Regression

Logistic regression is another technique that has shown success in fraud detection. According to (Alenzi and Aljehane, 2020), some of its advantages include its suitability for classification of widely distributed data, ease of implementation, efficiency in classifying unknown data as well as the fact that it can be easily extended to multiple classes.

4. DESIGN SPECIFICATION

This section highlights the proposed model's framework. This underlines the phases undertaken in the creation of the models. Firstly, the credit card transaction dataset was ingested into the Python environment where it was processed and prepared to be suitable for use. Thereafter, feature selection was employed to ascertain the features in the dataset which were most relevant for modelling. Subsequently, two machine learning algorithms were applied to train the model. After training, the model was then tested and evaluated. The figure below is a pictorial representation of the entire framework.

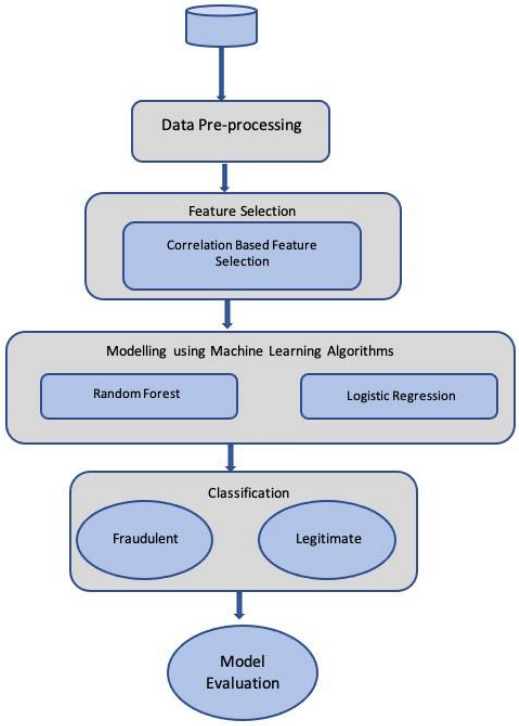


Figure 3: Framework for Proposed Credit Card Fraud Detection Model

4.1 System Requirement

This section tables both the software and Hardware requirements used for carrying out this research. These are highlighted in the table below.

Table 1:System Requirement

	Minimum Requirement	Current System Setup
Software	Chrome browser	Chrome browser
	Python Programming Language v3	Python Programming Language v3.8.3
	Anaconda Individual Edition v4	Anaconda Individual Edition v4.8.3
	Windows 8	Mac OS 11
Hardware	8GB RAM	16GB RAM
	150GB HDD	500GB HDD

4.2 Description and Functionality of the Mode

The model developed is a credit card fraud detection model that can classify credit card transactions into fraudulent and legitimate transactions, thus detecting credit card fraud. Here, the model was first trained using a subset of the dataset chosen for this research. It was then tested. Subsequently, evaluation was done to determine its accuracy as well as false negatives

i.e., number of fraudulent transactions being misclassified as legitimate. This is important as misclassification can ultimately lead to fraud which results in detrimental losses to any financial institution. Thus, the functionality of the model is to accurately classify credit card transactions into fraudulent and legitimate while achieving low false negatives.

5. IMPLEMENTATION

As stated in previous sections, 4 models were implemented. Therefore, this section highlights the implementation steps carried out as well as the tools used.

5.1 Tools and Language

The tools used for the implementation of the models include python programming language which gives access to several libraries for machine learning, data processing, data visualization amongst others.

Jupyter notebook assessed through Anaconda Navigator IDE was used to write the code.

5.2 Data Pre-processing

From the exploration, several steps were performed to adequately process the data before applying feature selection and machine learning algorithms. Firstly, columns with a high number of null values i.e., *merchant state, zip code,* and *errors* were deleted. Then the label column *Is Fraud?* was renamed to *Fraud*. After that, the $ symbol from the *amount* column was removed and converted from the object to float data type. Subsequently, object columns like *time, use chip, merchant name,* and *fraud columns* were encoded. Then the dataset was balanced using random under-sampling on the majority class (non-fraud) and oversampling on the minority class (fraud). Thus, the non-fraud class was reduced to one million records while the fraud class was increased from twenty-nine thousand (29,000) to one million (1,000,000). Finally, the *amount* was scaled using Robust Scaler.

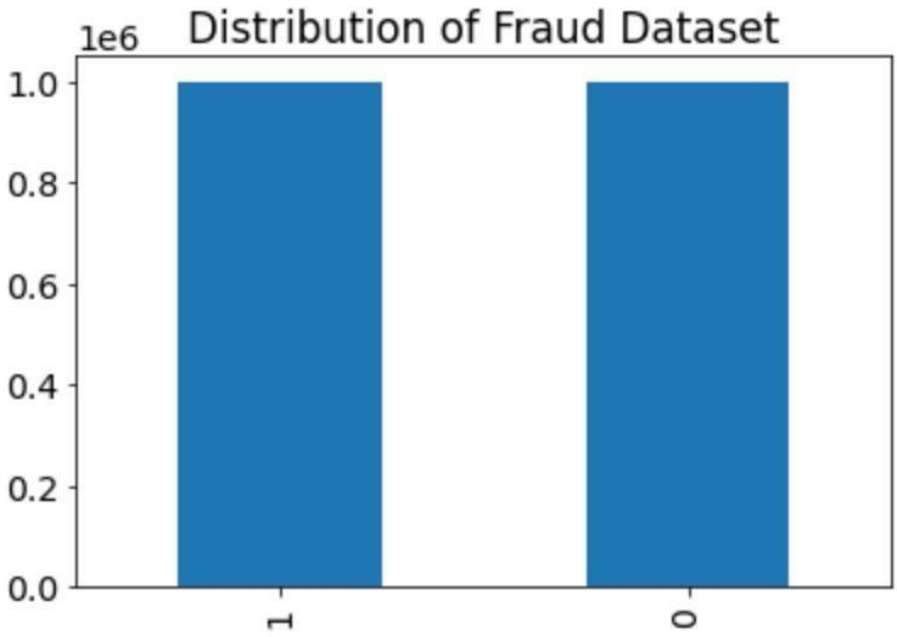


Figure 4: Dataset Distribution

5.3 Feature Selection

Here the correlation feature selection technique was employed to determine features that had a high correlation with the label column. The output is a correlation matrix which is scaled with numbers and colors depicting correlation strength. Thus, from the matrix, columns *Month, Card, Year, Day, User* were identified as having the least correlation with the label column and hence, were removed. Consequently, only 7 features were used to train the models i.e., the *fraud* column and 6 other independent variables. After this, the dataset was split into 80:20 for training and testing respectively.

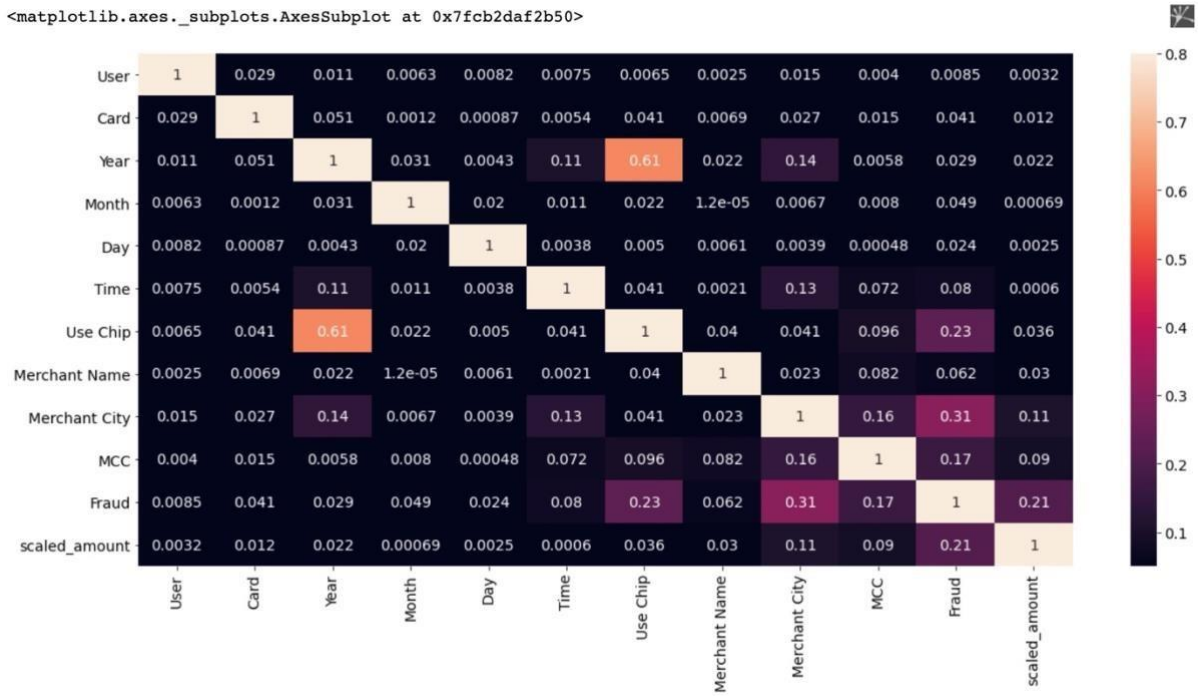


Figure 5: Correlation Matrix

5.4 Models Developed

Four models based on two machine learning algorithms were developed. Each machine learning model was implemented with and without feature selection. Their implementation is highlighted below:

5.4.1 Implementing Random Forest Classifier with/without Feature Selection

In implementing the random forest classifier, one was done with feature selection i.e., 7 features were used and the other without feature selection i.e., 12 features were used. Both involved the same steps. The default RandomForest Classifier was used. Therefore, the *RandomForestClassifier* library was imported from *sklearn.ensemble* package. Then the default *n_estimator* of 100, max_features = "auto" was used in creating the model. A random state was assigned to each model to ensure reproducibility whenever the program is run. The *fit()* method is then used to fit the classifier on the training dataset while the *predict()* method is used to test the model on the testing dataset.

5.4.2 Implementing Logistic Regression Classifier with/without Feature Selection

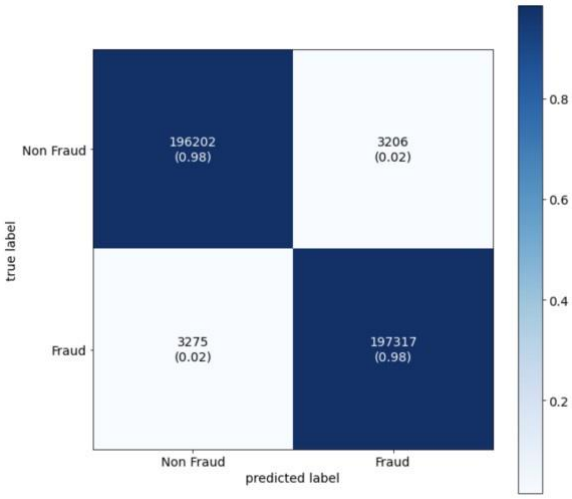
In implementing the Logistic Regression classifier, one was done with feature selection i.e., 7 features were used and the other without feature selection i.e., 12 features were used. Both involve the same steps. The logistic regression was imported from the *sklearn.linear_model* library. The *fit()* method is then used to fit the classifier on the training dataset while the *predict()* method is used to test the model on the testing dataset.

6. EVALUATION

This section presents the evaluation of all models developed in this research. The confusion matrix was used to ascertain the true positive, true negative, false negative, and false positive. The False-positive represents the number of legitimate transactions classified as fraudulent while the False-negative represents the number of fraudulent transactions classified as legitimate (Monedero et al., 2012). The True Positive refers to the number of fraudulent transactions accurately classified while the True negative refers to the number of legitimate transactions classified correctly. Other metrics such as recall, precision, Area Under Curve (AUC), and f1 score were also determined.

6.1 Evaluation of Random Forest Classifier with Feature Selection

This model achieved 98% accuracy. It has a similar false positive and false negative rate of 2%. This means there is a 2% chance of a credit fraud transaction being classified as a legitimate transaction. The AUC score shows the relationship between the false-positive and false-negative rates. The diagonal line on the AUC chart serves as a baseline of a model with a score of 50%. The farther the AUC curve is from the diagonal line, the better the model.



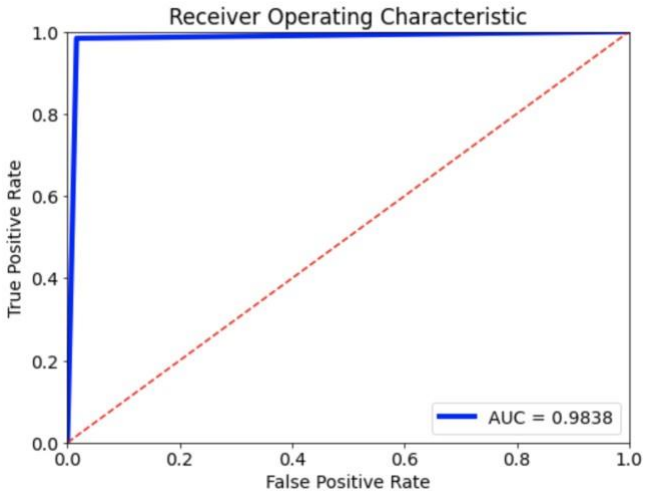


Figure 6:Confusion Matrix for RFC with Feature selection

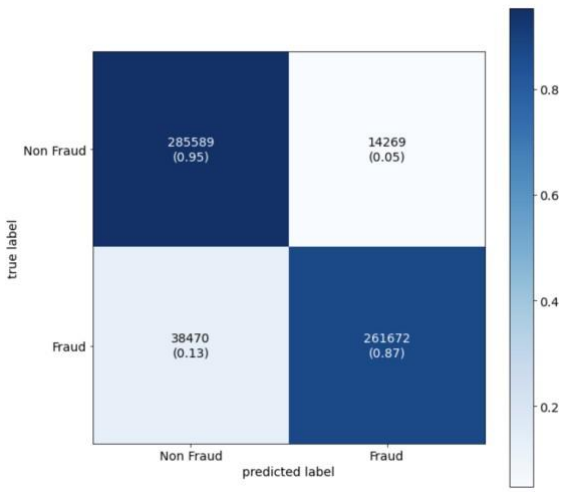
Figure 7:AUC ROC for RFC with Feature Selection

	precision	recall	f1-score	support
Non Fraud	0.9836	0.9839	0.9838	199408
Fraud	0.9840	0.9837	0.9838	200592
accuracy			0.9838	400000
macro avg	0.9838	0.9838	0.9838	400000
weighted avg	0.9838	0.9838	0.9838	400000

Figure 8: Classification Report for RFC with Feature selection

6.2 Evaluation Random Forest Classifier without Feature Selection

The random forest classifier model without feature selection performed a little less than the model with feature selection. It achieved an overall accuracy of 91% with a higher false positive and false negative value of 5% and 13% respectively. This higher false-negative value of 13% is a lot higher compared to the 2% gotten from the previous model with feature selection. Thus, making it less suitable for real-life application.



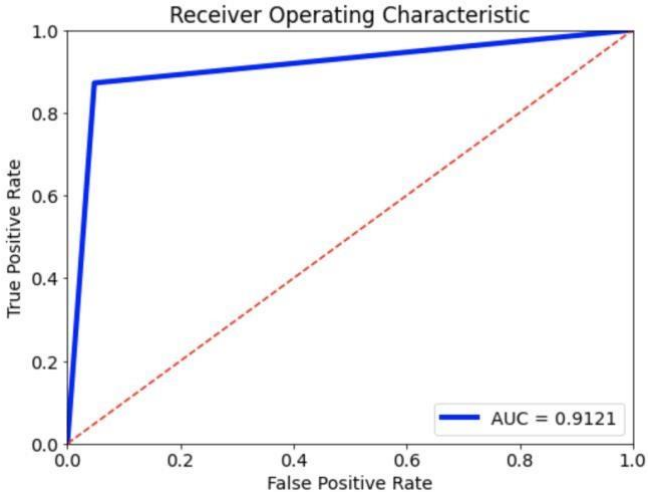


Figure 9:Confusion Matrix for RFC without Feature Selection

Figure 10: AUC ROC for RFC without Feature Selection

	precision	recall	f1-score	support
Non Fraud	0.8813	0.9524	0.9155	299858
Fraud	0.9483	0.8718	0.9085	300142
accuracy			0.9121	600000
macro avg	0.9148	0.9121	0.9120	600000
weighted avg	0.9148	0.9121	0.9120	600000

Figure 11: Classification Report for RFC without Feature Selection

6.3 Evaluation of Logistic Regression Classifier with Feature Selection

The logistic regression model with feature selection performed less than the random forest classifier. The model achieved an accuracy of approximately 70%. Similarly, the AUC score was also poorer than that of the random forest classifier. The model also had a higher false positive and false negative rate of 30% and 29% respectively. The false positives represent transactions that are legitimate but are wrongly classified as fraud while false negatives represent transactions that are fraudulent but are incorrectly classified as legitimate. A higher false-negative rate in this case would be more harmful.

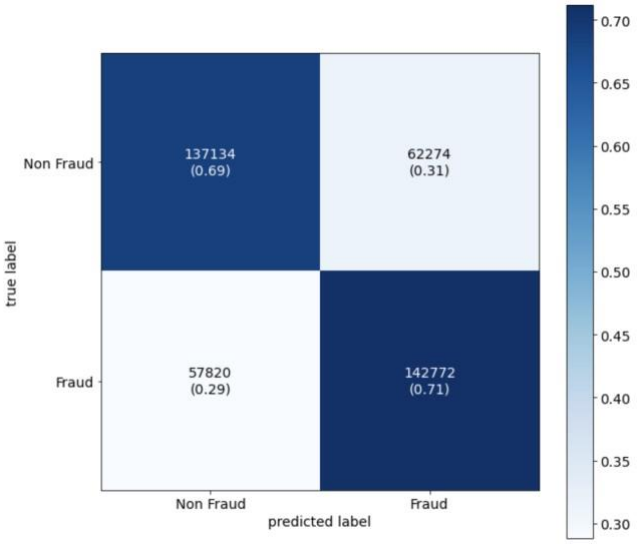


Figure 12:Confusion Matrix for LRC with Feature selection

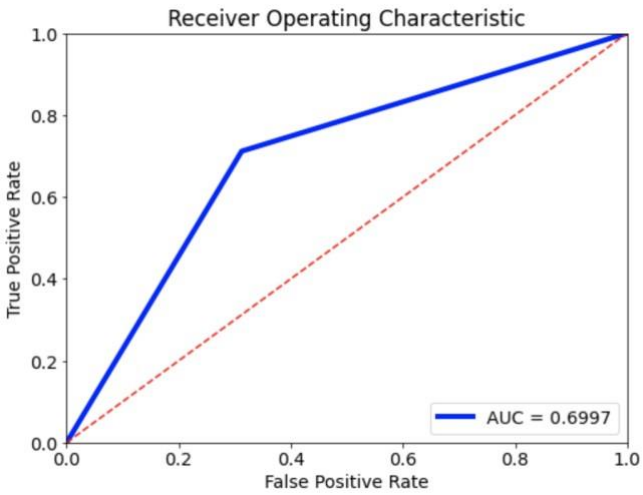


Figure 13:AUC ROC for LRC with Feature selection

	precision	recall	f1-score	support
Non Fraud	0.7034	0.6877	0.6955	199408
Fraud	0.6963	0.7118	0.7039	200592
accuracy			0.6998	400000
macro avg	0.6999	0.6997	0.6997	400000
weighted avg	0.6998	0.6998	0.6997	400000

Figure 14: Classification Report for LRC with Feature selection

6.4 Evaluation of Logistic Regression Classifier without Feature Selection

Like the comparison between the 2 random forest models, the logistic regression model without feature selection also performed less than the model with feature selection. It achieved an overall accuracy of 66%, making it the least suitable of the 4 models generated during this research.

	precision	recall	f1-score	support
Non Fraud	0.6652	0.6341	0.6493	299858
Fraud	0.6508	0.6811	0.6656	300142
accuracy			0.6576	600000
macro avg	0.6580	0.6576	0.6574	600000
weighted avg	0.6580	0.6576	0.6574	600000

Figure 15:Classification Report for LRC without Feature Selection

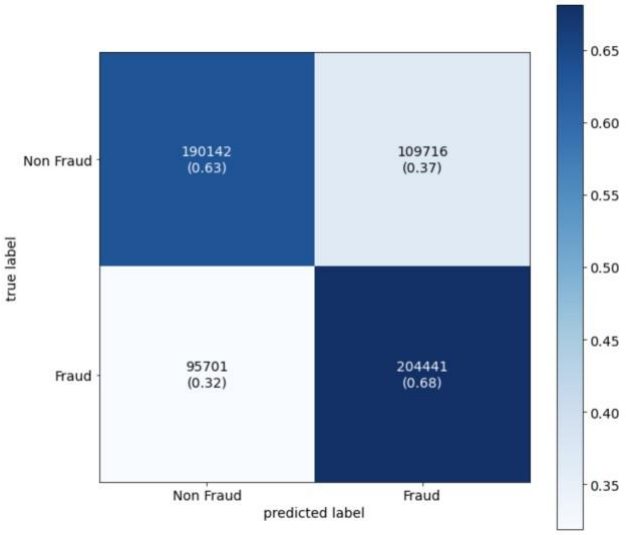


Figure 16: Confusion Matrix for LRC without Feature Selection

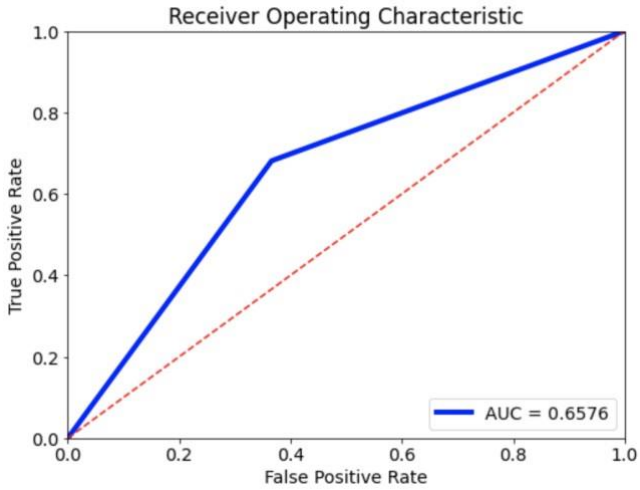


Figure 17:AUC ROC for LRC without Feature Selection

7. CONCLUSION AND FUTURE WORK

This paper was aimed at answering the Research Question: **How accurate are machine learning algorithms using feature selection techniques for detecting credit card fraud?** In achieving its objectives, several credit card fraud detection classifier models were developed using machine learning algorithms with and without feature selection. This was to compare if using feature selection improves the accuracy of the models. After implementation, all models were evaluated. The results of the experiments showed that the models with feature selection achieved higher accuracy when compared to their corresponding machine learning models without feature selection. Overall, the classifier based on Random Forest with feature selection achieved the highest accuracy of 98% and the best F1 score which shows how precise the model is.

Consequently, the random forest classifier with feature selection is suitable to be used in the credit card fraud detection model as it achieves high accuracy with a good F1 score as well as the lowest false negative.

Furthermore, this research was limited by the lack of a range of credit card transaction records available online. Also, another limitation was the high computing power and long training times involved in machine learning operations. Future work could consider implementing other machine learning techniques and deep learning techniques against the dataset used in this research.

REFERENCES

Altman, E. R. (2019) 'Synthesizing Credit Card Transactions', *arXiv:1910.03033 [cs]*. Available at: http://arxiv.org/abs/1910.03033

Bahnsen, A., Aouada, D., and Ottersten, B. (2015) 'Example-dependent cost-sensitive decision trees', Expert Systems with Applications, 42(19), pp. 6609–6619. doi: 10.1016/j.eswa.2015.04.042.

Bentley et al (1997) 'Hierachical Crossover in Genetic Algorithm'

Carcillo, F. et al. (2019) 'Combining unsupervised and supervised learning in credit card fraud detection', Information sciences.

Emura, T., Matsui, S. and Chen, H.-Y. (2019) 'compound. Cox: univariate feature selection and compound covariate for predicting survival', Computer methods and programs in biomedicine, 168, pp. 21–37.

Ezawa et al (1996) 'Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management'. In Proceedings of the 13th International Conference on Machine Learning.

G. Louppe, "Understanding Random Forests," Cornell Univ. Libr., 2014.

J.R Dorronsoro et al (1997)" Neural fraud detection in credit card operations"
James, G. et al. (2013) An Introduction to Statistical Learning. New York, NY: Springer New York (Springer Texts in Statistics). doi: 10.1007/978-1-4614-7138-7.

Karegowda, A. G., Manjunath, A. S. and Jayaram, M. A. (2010) 'Comparative study of attribute selection using gain ratio and correlation based feature selection', International Journal of Information Technology and Knowledge Management, 2(2), pp. 271–277.

M. A. Hall, "Correlation-based Feature Selection for Machine Learning," 1999.

M. Suresh Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit Card Fraud D etection Using Random Forest Algorithm," *IEEE,* 2019.

Maes, S., Tuyls, K., Vanschoenwinkel, B. & B Manderick. 2002. Credit Card Fraud Detection using Bayesian and Neural Networks, Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies.

Monedero, I. et al. (2012) 'Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees', International Journal of Electrical Power & Energy Systems, 34(1), pp. 90–98.

Patidar, R. and Sharma, L. (2011) 'Credit Card Fraud Detection Using Neural Network', in In IJSCE ISSN: 2231-2307, Volume-1, Issue-NCAI2011.

Pozzolo, A. D. (2015) 'Adaptive Machine Learning for Credit Card Fraud Detection', Université Libre de Bruxelles, p. 199.

Pun, J. K.-F. (2011) Improving credit card fraud detection using a meta-learning strategy.
R. Duangsoithong and T. Windeatt, "Correlation-based and causal feature selection analysis for ensemble classifiers," 2010, doi: 10.1007/978-3-642-12159-3_3.

Raj, S. B. E. and Portia, A. A. (2011) 'Analysis on credit card fraud detection methods', in 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET). IEEE, pp. 152–156.

Sahin and Duman (2011) 'Detecting credit card fraud by ANN and logistic regression'

Samaneh, S. et al. (2016) 'A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective', arXiv:1611.06439 [cs]. Available at: http://arxiv.org/abs/1611.06439 (Accessed: 7 August 2021).

Shiyang Xuan, Guanjun Lium Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang, "Random Forest for Credit Card Fraud Detection," IEEE, 2018.

Suman (2014) "Survey Paper on Credit Card Fraud Detection" , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "The KDD Process forExtracting Useful Knowledge from Volumes of Data," Commun. ACM, vol. 39, no. 11,pp. 27–34, Nov. 1996, Accessed: Nov. 21, 2020. [Online]. Available:https://sceweb.uhcl.edu/boetticher/ML_DataMining/p27-fayyad.pdf.

Van Vlasselaer, V. et al. (2015) 'APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions', Decision Support Systems, 75, pp. 38–48. doi: 10.1016/j.dss.2015.04.013.

Wen-Fang YU and Na Wang (2009) "Research on Credit Card Fraud Detection Model Based on Distance Sum" International Joint Conference on Artificial Intelligence.

Yan, K. and Zhang, D. (2015) 'Feature selection and analysis on correlated gas sensor data with recursive feature elimination', Sensors and Actuators B: Chemical, 212, pp. 353–363.

Yvan Lucas, et al. (2019) 'Credit card fraud detection using machine learning: A survey'