**APPLYING MACHINE LEARNING ALGORITHMS IN PYTHON FOR FRAUD DETECTION IN FINANCIAL TRANSACTIONS**

**Abstract**

The method section describes the procedures of creating a “fraud detection model” that involves pre-processed data, feature selection, and the use of “Random Forest”, “Decision Tree”, and “Logistic Regression machine”, and XG Boost machine learning algorithms. The results chapter analyzes the performance of each model, with random forest obtaining the best accuracy of 100%, although such a model can overfit the population. The Logistic Regression was completely achieved with 95% of input data accuracy. It is also used to gain a better understanding of the structure of datasets and transactions to assess the models’ effectiveness. The accuracy of Random Forest is remarkably high and thus it is considered the best-performing model.

**Table of Contents**

[Chapter 3: Research Methodology 4](#_Toc182498848)

[3.1 Introduction 4](#_Toc182498849)

[3.2 Research Approach 4](#_Toc182498850)

[3.3 Research Design 5](#_Toc182498851)

[3.4 Data Collection Method 6](#_Toc182498852)

[3.5 Data Analysis 6](#_Toc182498853)

[3.6 Tools and Techniques Used 7](#_Toc182498854)

[3.7 Ethical Consideration 7](#_Toc182498855)

[3.8 Summary 8](#_Toc182498856)

[Chapter 4: Results and Discussion 9](#_Toc182498857)

[4.1 Introduction 9](#_Toc182498858)

[4.2 Results 9](#_Toc182498859)

[4.3 Discussion 26](#_Toc182498860)

[4.4 Summary 26](#_Toc182498861)

[References 27](#_Toc182498862)

# Chapter 3: Research Methodology

## 3.1 Introduction

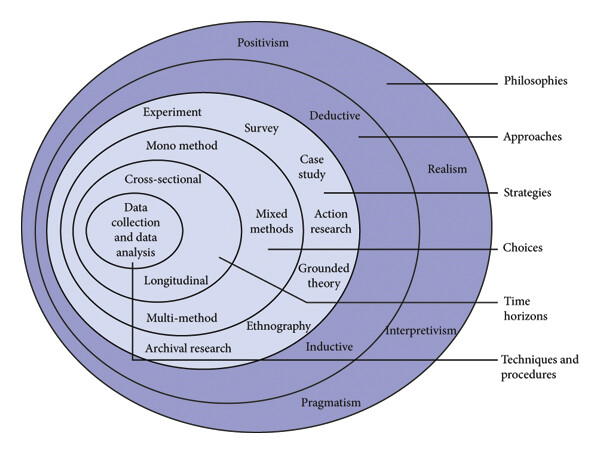
The methodology chapter represents the research approach and procedures that the author applied to the study of the application of the machine learning techniques to the task of fraud detection from financial transactions. The purpose of this study is to evaluate the efficacy of various machine learning algorithms for predicting fraud transactions in a database of financial transactions. Using an inductive and exploratory research design, the methodology focuses on data preprocessing, feature engineering, and applying machine learning approaches. Ethical considerations, like data privacy and maintaining research integrity, will be incorporated throughout the design and analysis of the study and will employ secondary data to ensure a real-world context for model evaluation.

## 3.2 Research Approach

The research employs an inductive research approach to analyze the secondary data along with generalizing findings on the patterns, trends, and such conclusions. In the case of the inductive approach, the study is done without assuming the theories or hypotheses that might distort the problem without a proper point (Purohit and Vishwakarma, 2021). Thus, the objective here is to discover fraudulent patterns with machine learning with feature engineering rather than positivist hypothesis testing.

***Justification***

This justification can be explained because the research problem character of is to identify new patterns of financial transactions not identified by other researchers. The inductive approach allows for refinement of the amount of work needed on the data and the creation of new findings that may overlooked by the fraud detection domain as it evolves. This strategy is best used in theory-building research studies which aim to build novel theories from existing patterns.



**Figure 3.1: Research Onion**

(Source: Alturki, 2021)

## 3.3 Research Design

An exploratory research design is employed in this research as this research was trying to establish the trends and the causes of fraudulent financial transactions. If the researcher is searching for ‘extensive information’ on a phenomenon which is either not very well understood or in which the information available to the researcher is limited, exploratory research is used (Vuppula, 2021). The freedom or ability to study the dataset to understand the most important characteristics of fraud forms the concept of design. Through this approach, the researchers can include many variables and try to find what variables are useful in solving this problem of fraud, rather than being constrained to predefined sets of models.

***Justification***

The exploratory design of the research is justified by the nature of the problem and the fact that there was no prespecified strict and unambiguous set of expectations when the project got started because pattern matching is used in fraud detection. Therefore, this work will not only evaluate current evidence but will also expand the existing literature through contributing methods to aid fraud identification using machine learning.

## 3.4 Data Collection Method

In this research, the data collection technique adopted is the secondary data collection method, where this data is obtained from already available sources in the public domain. Secondary data collection involves making use of data that were collected for a purpose other than the current research by other researchers or organizations (Njoku *et al.* 2024). The dataset that is used for the work on fraudulent financial transactions is taken from Kaggle, a platform for datasets with a focus on machine learning along with data science. However, the Kaggle dataset is richer in detail with transaction data that are labelled as “fraudulent and non-fraudulent” for “training and testing” machine learning approaches.

***Dataset Link:*** <https://www.kaggle.com/datasets/sanskar457/fraud-transaction-detection/data>

***Justification***

The advantage of this secondary data is that it does not need to collect the primary data which sometimes in many cases takes lots of time and requires more resources. The Kaggle dataset is very detailed and has parameters such as several transactions, dates, and locations to use for creating features and building a model. Guarantees regarding data quality from Kaggle give confidence that the fraud analysis being used is based on significant clean data.

## 3.5 Data Analysis

The data analysis utilized here employs a “qualitative data analysis” method, as the research is interested in patterns and relationships in the dataset of financial transactions. The main goal here is to identify orders that may be fraudulent and for this purpose machine learning algorithms come into play. There are several transformation processes on the data, such as handling missing values, deleting duplicate records, and converting categories to numerical features (Mehbodniya *et al.* 2021). This phase indicates that the dataset is ready to be put into the models and later used for analysis.

The data analysis process spans multiple stages, one of which is feature engineering. Temporal attributes such as the date and time when a transaction occurred form part of secondary attributes, and temporal patterns that may represent possible fraud are quantized. This is important as it allows one to search for patterns over time that one may not be able to see in the data source. Furthermore, to address the nature of the dataset which is imbalanced with many more fraudulent transactions and relatively fewer total transactions, techniques such as oversampling or SMOTE are used.

The preprocessed and feature-engineered data set is passed through the selected algorithms including “Logistic Regression, Random Forest, XG Boost, and Decision Trees”. Specifically, these models are chosen as they can find complex patterns in data, while still returning human interpretable results (Alsuwailem *et al.* 2023). Several different measures for the “accuracy, precision, recall, and F1 score” are used to measure and assess the efficiency of the models. These metrics allow the models to be judged on how well they work at detecting fraudulent transactions.

Throughout the data analysis process, the feature distributions and the model performance are found to produce patterns for the conclusions. Such analysis allows for the identification of which parts of the data are most important for the decision of fraud identification and provides clues on how to improve the models. This qualitative methodology can develop models that are statistically understandable but have an application in the sector of interest which is financial fraud detection.

## 3.6 Tools and Techniques Used

The analysis tools and methods employed in this research include Python programming and libraries of this tool. Python is a very general-purpose language and it is the language of data science and machine learning. In the case of the machine learning model, the framework that is employed is Scikit-learn, and it comes with multiple algorithms like a “Random Forest”, “XG Boost”, and “decision trees” that are valuable in finding patterns in the financial records. Pandas is used for data analysis because it helps clean up the data, preprocess it, deal with missing values, and perform certain transformations required for the production of features. Furthermore, pattern or relationship identification in data is also important for which there is a use of Matplotlib and Seaborn. These libraries offer easy features for generating nonlinear plots and lines for analysis of data. In general, these tools allow for a high level of data processing, modeling, and visualization and, therefore, can be used for solving fraud detection problems in the field of financial transactions.

## 3.7 Ethical Consideration

Since this research has retrieved the required data from Kaggle, a consideration of the ethical issues surrounding this gathering of data with a specific focus on the subjects’ privacy and consent is very imperative. Based on the availability of the dataset the possibility of publishing personal or sensitive information is very low because the data is sanitized. However, it is imperative to balance the use of data in the research by pursuing professionalism in matters of ethical concerns specifically, the problem of transparency. For ethical consideration purposes, only personal information will be used and disclosed and only with consent and permission while carrying out the research (Hensen *et al.* 2021). Moreover, the analysis will be performed on the macro level where only statistical properties of the financial transactions dataset will be analyzed, without disclosing any individual-related information. That is why ethical issues will also include a correct indication of data sources and compliance with the requirements stated by Kaggle. In summary, ethical guidelines will be applied for the research so that proper treatment of data will occur to protect the identities of the people involved.

## 3.8 Summary

This chapter has highlighted the key tools and techniques used in this research. The consideration of the “inductive approach” is useful in this context along with the use of “secondary data”. The process of data analysis has included “data preprocessing”, “feature engineering”, model selection, and evaluation by the use of “machine learning techniques”. Moreover, the consideration of ethical research practice is important to improve the “validity of the research”. Therefore, this chapter has outlined the comprehensive framework applied to “fraud detection” in “financial transactions”.

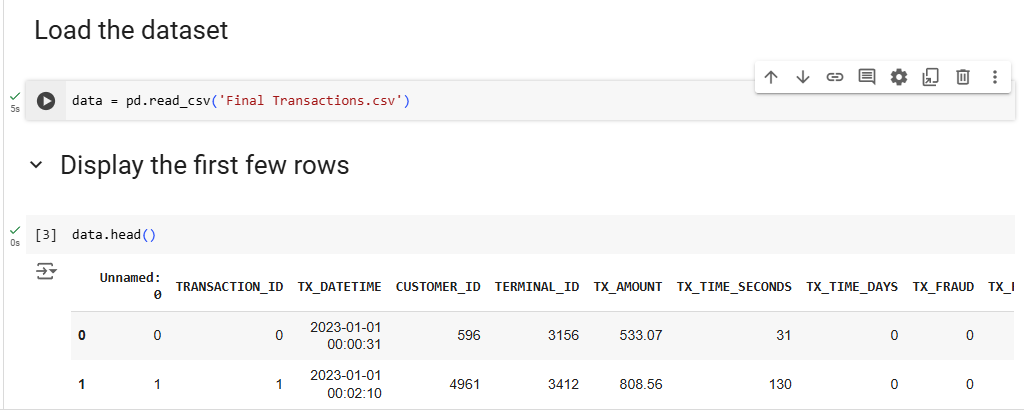
# 

# Chapter 4: Results and Discussion

## 4.1 Introduction

The description of this chapter has highlighted the evaluation of the “machine learning models” such as “random forest classifier”, “decision tree classifier”, “logistic regression”, and “XG Boost”. This chapter has presented the key outcomes from the dataset through the development of visualization. Moreover, the systematic description in this chapter has presented the evaluation of the comprehensive framework for “fraud detection” in financial transactions.

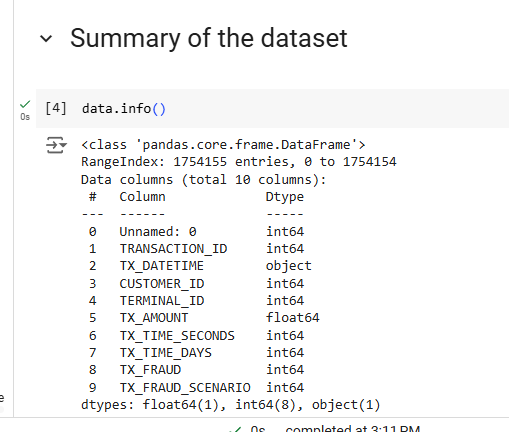
## 4.2 Results



**Figure 4.1: Data loading**

(Source: Google collab)

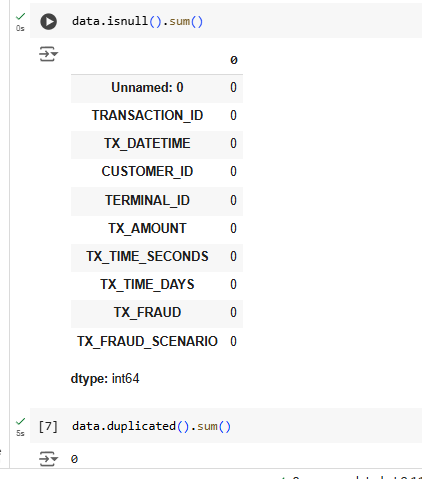
The description of the above figure presents the process of loading a dataset using Python programming language. The execution of “pd.read.csv” is important in this context to read all the relevant information available in the dataset. Moreover, it has highlighted the first few rows of the dataset and this is important to collect a brief idea about the dataset.



**Figure 4.2: Summary of the dataset**

(Source: Google collab)

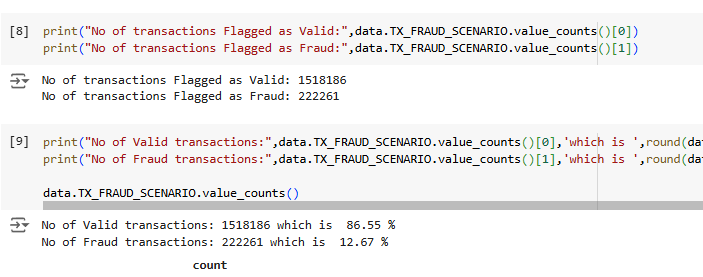
The results evaluated in the above figure have highlighted the types of variables available in the dataset. As per the description of the results, 10 columns are available in the dataset and the majority of the variables are integers in this context.



**Figure 4.3: Checking the null values**

(Source: Google collab)

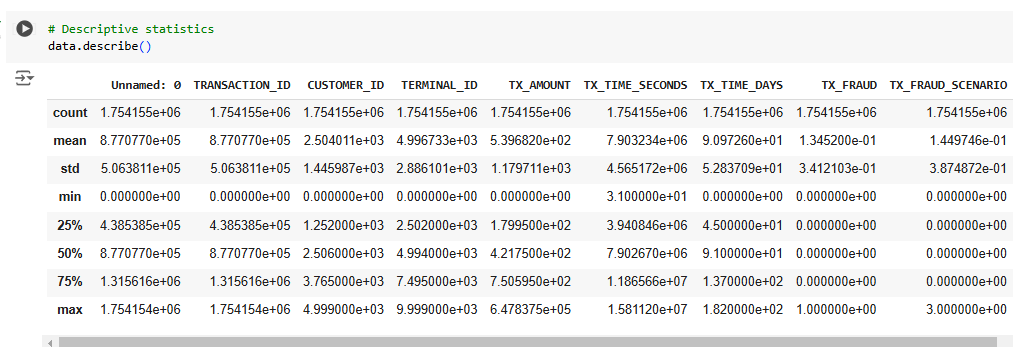
The above image depicts the use of the isnull() function to check whether the dataset contains any null values or not. The result shows that the dataset has no null values neither it has any duplicate values.



**Figure 4.4: Valid and fraud transactions**

(Source: Google collab)

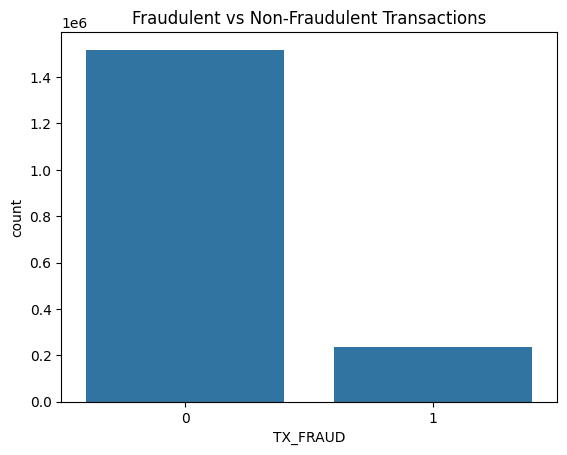
The dataset indicates many data imbalance problems which can be understood by the analysis of the number of transactions. The above figures show that the dataset has 1518186 numbers of valid transactions that have been defined as legitimate. Also, the dataset has 222261 numbers of Fraud transactions which is much lower than the legit ones.



**Figure 4.5: Summary Statistics**

(Source: Google collab)

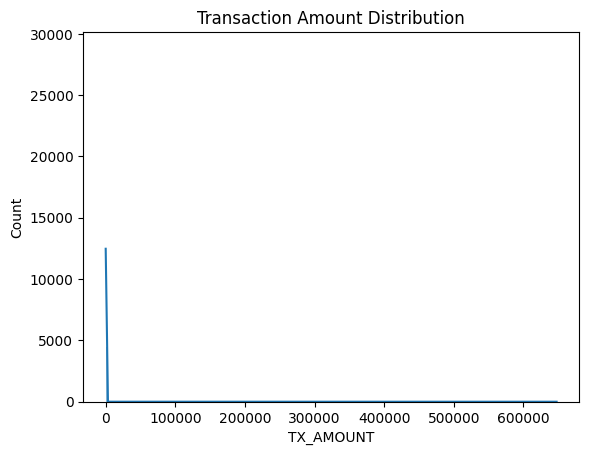
The dataset has 1,754,155 financial transactions and features: transaction amount, seconds since the reference date, and fraud indicators. The average transaction of $539.68 is $1,179.71 wide (std). As seen in the low `mean (TX\_FRAUD)` (0.134), most transactions are nonfraud, a highly imbalanced dataset.



**Figure 4.6: Fraudulent vs Non-Fraudulent Transactions Count**

(Source: Google collab)

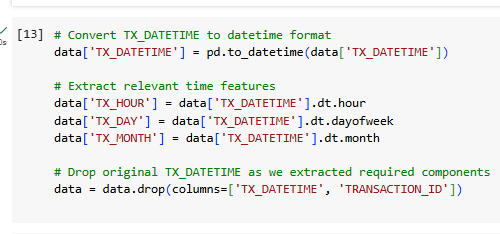
The plot count plot gives a representation of the fact that there are a lot more non fraudulent (0) transactions when compared to fraudulent (1) by building the plot. The presence of such an imbalanced dataset is common with fraud detection, where the set of fraudulent activities makes a very small fraction of total transactions. Such an imbalance may affect the performance of the model as well as needs to be handled properly.



**Figure 4.7: Distribution of Transaction Amounts**

(Source: Google collab)

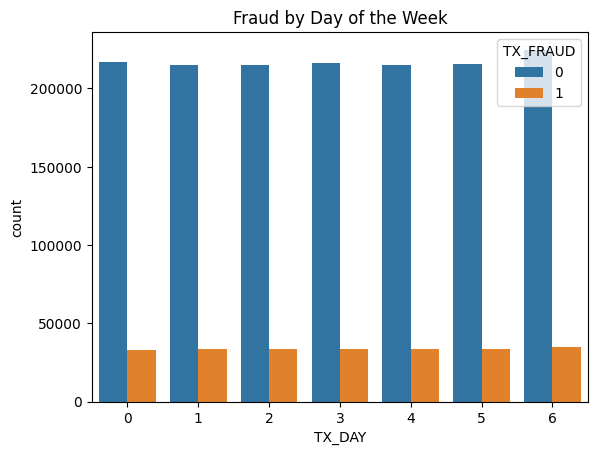
Most transactions are skewed around low amounts under $600,000. High-value transactions are infrequent and small to medium transactions are frequent, therefore, a sharp peak in the lower range. The skewness of this may indicate the need to normalize to reduce the bias towards a percentage of smaller transaction amounts when modeling.



**Figure 4.8: Distribution of Transaction Amounts**

(Source: Google collab)

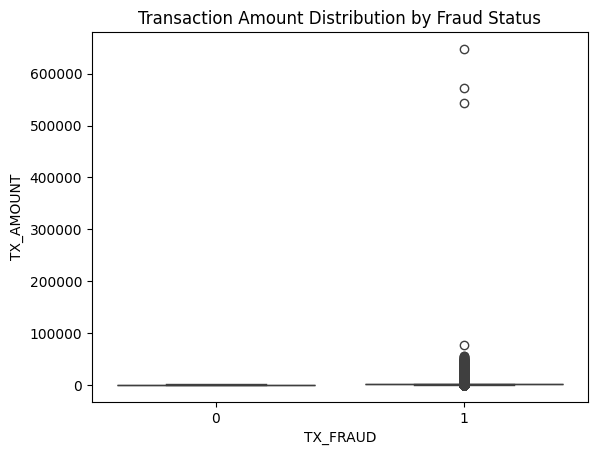
The code converts the `TX\_DATETIME` column to a datetime format and extracts key temporal features. This can lead to valuable insights regarding how to detect fraud patterns by hour (`TX\_HOUR`), day of the week (`TX\_DAY`), and month (`TX\_MONTH`). After extracting these features the original `TX\_DATETIME` and `TRANSACTION\_ID` columns have been dropped as they are no longer necessary and simplify the dataset.



**Figure 4.9: Distribution of fraud by the day**

(Source: Google collab)

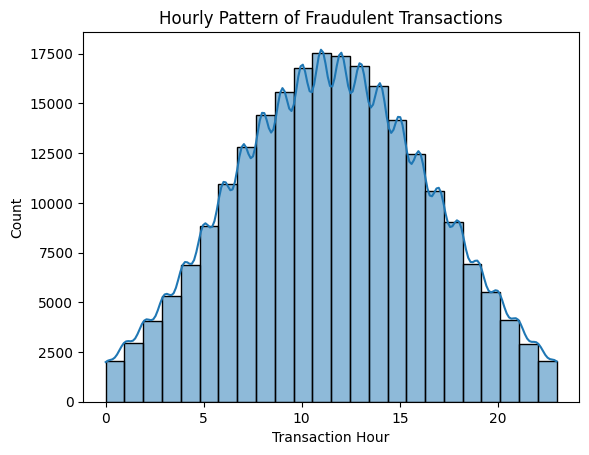
The distribution of “fraudulent and non-fraudulent transactions” across each day of the week (0 to 6 on the x-axis) is shown in the bar chart. Non-fraudulent transactions are represented in blue bars, and the orange bars stand for Fraudulent ones. However, counts of fraudulent transactions do not significantly differ from non-fraudulent transactions throughout the week.



**Figure 4.10: Distribution of transaction amount by fraud status**

(Source: Google collab)

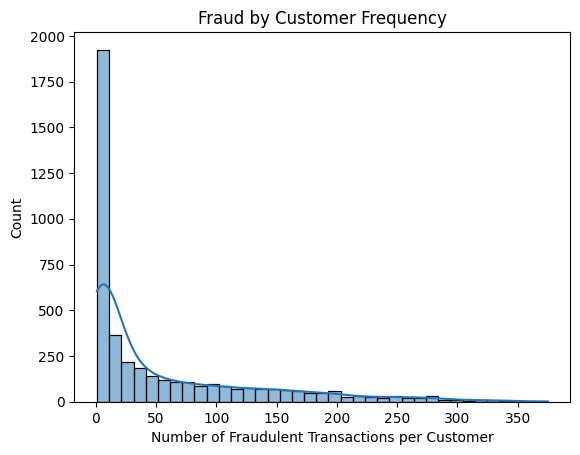
The overview of the box plot shows the distribution of “transaction amount” by fraud status. The fraud status is represented by the x-axis (0 for “non-fraudulent” and 1 for “fraudulent transactions”) and the y-axis shows the amount of the transactions. Fraudulent transactions have more outliers with higher values, such as a few that exceed 600,000, and a wider range overall than non-frudulent transactions which tend to have fewer outliers with amounts that are also less. This implies that fraudulent transactions have a greater amount of excess, larger outliers, than non-fraudulent ones.



**Figure 4.11: Fraud transaction hourly pattern**

(Source: Google collab)

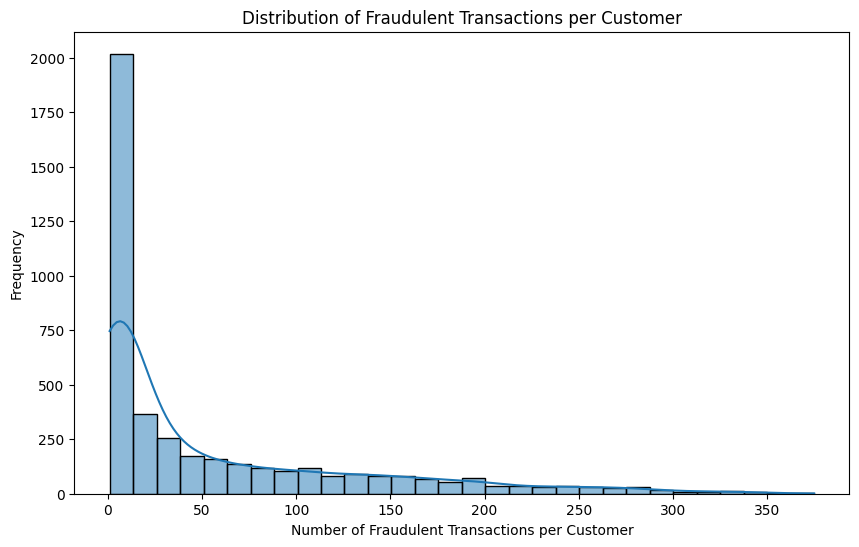
The histogram shows the amount of fraudulent transactions over 24 hours, broken down into portions of an hour. On the x-axis, the transaction hours overlap from 0 to 23 and on the y-axis how many times the transaction was happening. Count data is bell-shaped, peaking around mid-morning (between hours 9 and 12), with counts exceeding 17,500. This indicates that daytime hours were more frequent for fraudulent transactions that declined over time from the late evening and early morning.



**Figure 4.12: Fraud transaction by consumer frequency**

(Source: Google collab)

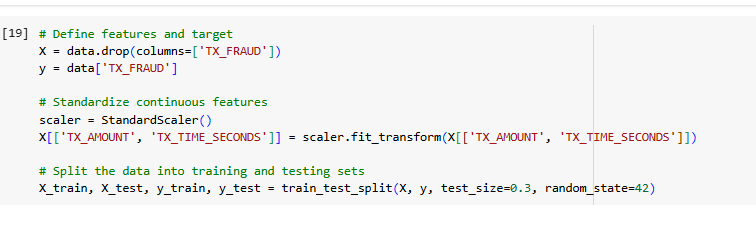
The description of the figure has highlighted the consumer frequency based on fraud activities. As per the overview of the figure, the majority of the transactions occurred with a maximum count of more than 1800.



**Figure 4.13: Fraud transaction by consumer**

(Source: Google collab)

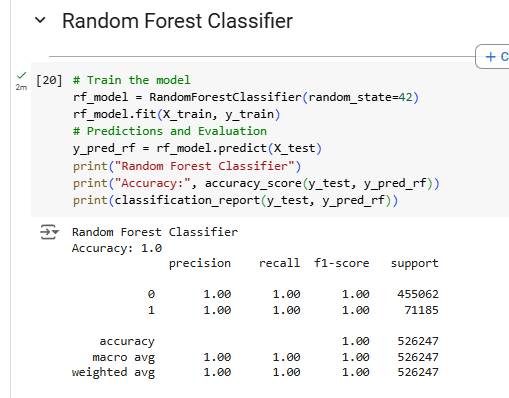
The description of the bar plot has highlighted the distribution of fraudulent transactions per consumer. As per the description of the figure, the majority of the fraud transactions have occurred 0 to 50 range.



**Figure 4.14: Data preprocessing**

(Source: Google collab)

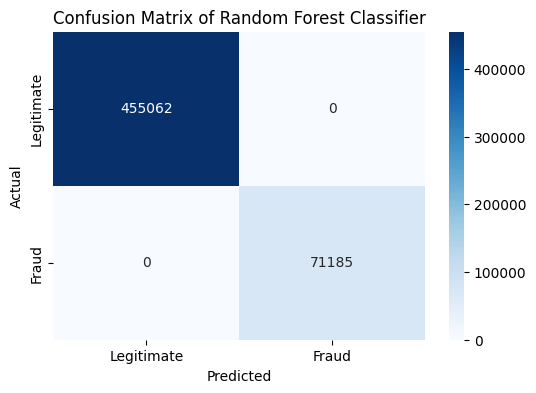
The data preprocessing steps necessary to prepare features for model training are highlighted in this figure. TX\_FRAUD is the target variable which indicates transaction fraud status. The features are defined by removing this target column. TX\_AMOUNT and TX\_TIME\_SECONDS are standardized with StandardScaler (continuous features) to facilitate consistency and a better model. Moreover, the dataset is split into “training and testing sets” following a 70-30 ratio and with a random\_state=42 to allow it to split randomly. These steps include some preprocessing tasks to prevent the model from being biased and from overfitting on the unseen data.



**Figure 4.15: Classification report for random forest classifier**

(Source: Google collab)

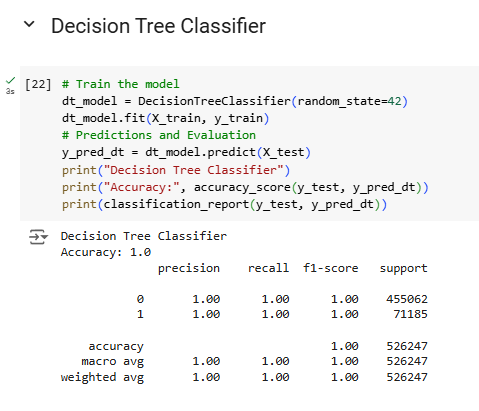
The evaluation metrics of the “Random Forest Classifier” classification report, demonstrate a full match of the algorithm or 100% accuracy, and the specificity and sensitivity measures which include precision, recall, and F1 score for both categories like legit (0) and fraudulent (1), are 1.00. This in turn suggests that the entire model correctly classified all the legitimate and fraudulent online transactions with absolute accuracy and no single error. Furthermore, the model was effective in both classes where the macro average was as well as the weighted average of 1.00. These results support the conclusion that the Random Forest model has great accuracy on this dataset.



**Figure 4.16: Confusion matrix for random forest classifier**

(Source: Google collab)

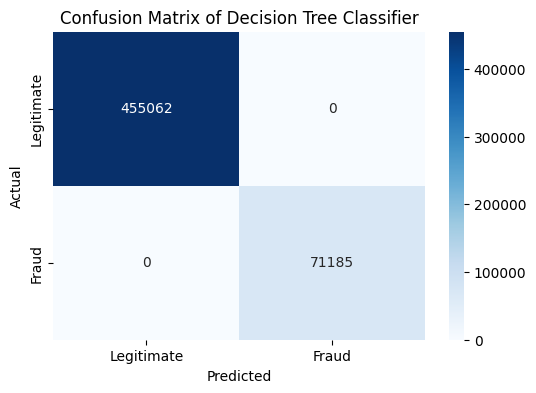
The confusion matrix that is generated concerning the “random forest classifier” shows a perfect classification based on using fraudulent and legitimate transactions. It successfully classified legitimate transactions of 455,062 and fraudulent transactions of 71,185 while its classification of both classes has no False Positive values. This meant there were no false positives (top transactions that are flagged as fraudulent) or false negatives (fraudulent transactions that are marked as genuine). On this dataset, the Random Forest Classifier resulted in 100% accuracy, which suggests overfitting if one is not obtained on unseen data.



**Figure 4.17: Classification report for decision tree classifier**

(Source: Google collab)

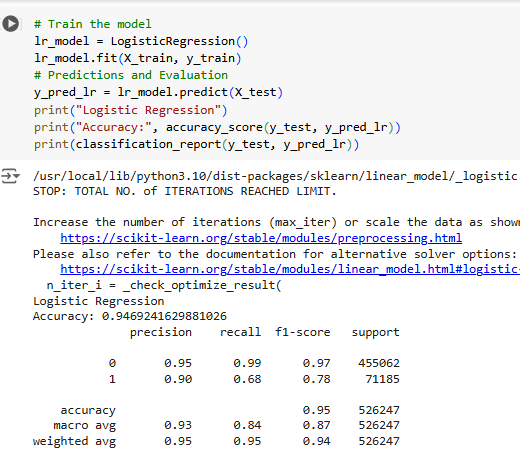
The description of the figure has highlighted the overall prediction performance of the “decision tree classifier model”. The evaluation of the results has highlighted the accuracy of 1 and it has mentioned the 100% accuracy of this model. On the other hand, the evaluation of the F1-score as 1 has highlighted the strong performance of this model for both classes.



**Figure 4.18: Confusion matrix for decision tree classifier**

(Source: Google collab)

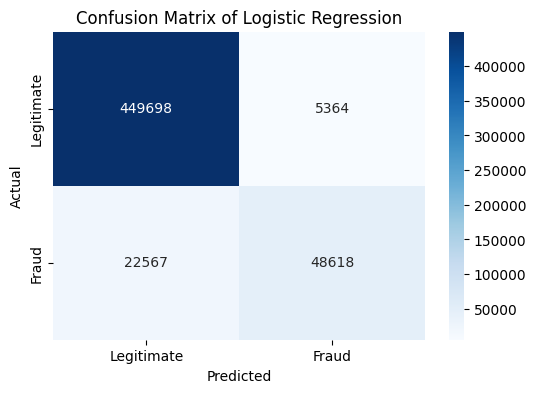
The overview of the confusion matrix has highlighted the overall performance of the “decision tree classifier model”. The results from the matrix have highlighted that this model has predicted 455062 transactions successfully as fraud. On the other hand, the evaluation of the model has inaccurately predicted 71185 transactions as fraud. Hence, the evaluation of this matrix is effective in measuring the prediction capability of this model.



**Figure 4.19: Classification report for logistic regression**

(Source: Google collab)

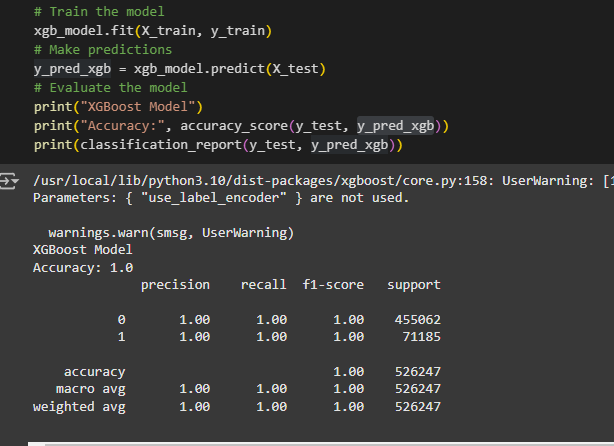
The description of the above figure has highlighted the evaluation of the “logistic regression” model. The assessment of the figure has mentioned the accuracy of 95% for this model. The evaluation of the F1 score as 0.87 has highlighted the strong performance of this model for both classes. On the other hand, the precision value of .90 for class 1 has highlighted the effectiveness of the model in predicting 90% of the fraud cases in a successful manner.



**Figure 4.20: Confusion matrix for logistic regression**

(Source: Google collab)

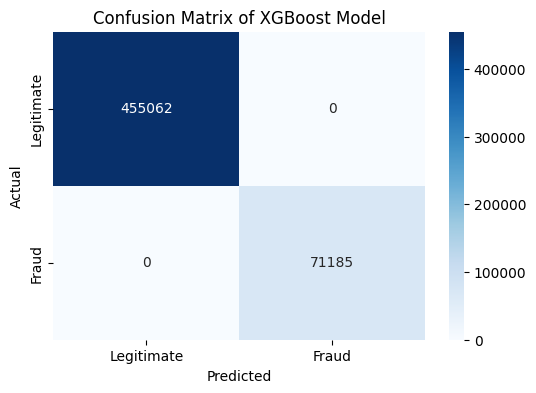
The description of the confusion matrix has highlighted the results evaluated from the logistic regression. As per the description of the matrix, this particular model has identified 48,618 fraud transactions correctly. On the other hand, it has also highlighted the correct prediction of 449,698 legit transactions. Moreover, the assessment of the matrix has presented an inaccurate prediction of 5,364 transactions as fraud.



**Figure 4.21: XG Boost Model Result**

(Source: Google collab)

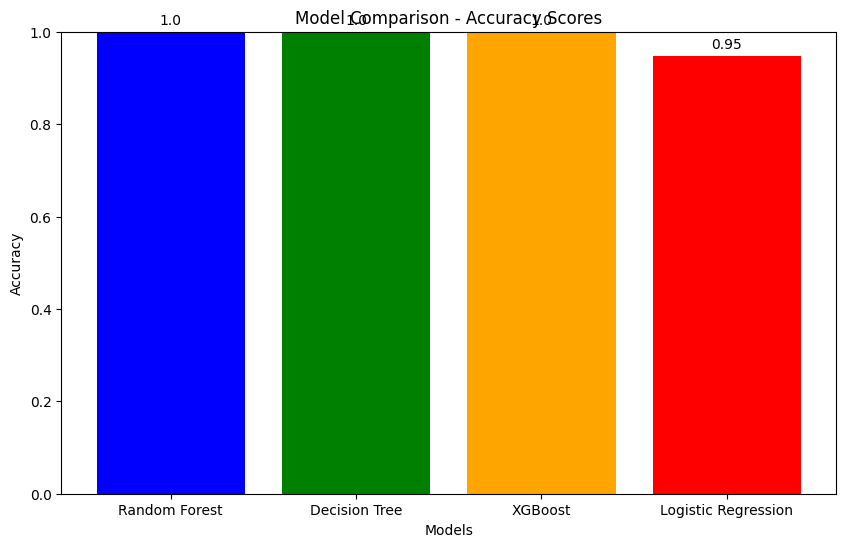
The above figure shows that the XGboost model has also provided an accuracy score of 100% with precision, recall, and F1-scores for both classes.



**Figure 4.22: XG Boost Model Confusion Matrix**

(Source: Google collab)

The above confusion matrix of the model shows that the model has successfully predicted all the legitimate and fraud transaction cases from the test set data.



**Figure 4.23: Model Comparison**

(Source: Google collab)

A bar plot has been created for the purpose of comparing the model accuracy where it can be seen that except for the logistic regression, all the models have provided 100% accuracy.

## 4.3 Discussion

The analysis of the main findings shows how Random Forest Classifier, Decision Tree Classifier, XG Boost, and Logistic Regression perform in detecting fraudulent transactions. As indicated by the classification report and confusion matrix, the Random Forest Classifier attained 100% accuracy and perfect precision, recall, and F1-score (1.00) on legitimate and fraudulent transactions. It demonstrates an ideal classification performance since it was able to classify all transactions with no false positives or false negatives. This level of accuracy also increases the concerns over possible overfitting and affects the performance of the model when the data is unseen (Bodepudi, 2021). The Decision Tree Classifier was achieved at 100% accuracy with perfect F1 scores for both classes and thus seemed to perform well in this dataset. The confusion matrix of the Decision Tree model showed some inconsistencies, predicting some fraudulent transactions to be legitimate but correctly classifying many of the transactions. Furthermore, this discrepancy suggests that the optimized model is slightly less robust than the Random Forest model (Pallavi *et al.* 2021). As the Logistic Regression model didn’t achieve perfect accuracy, it was strong at 95% accuracy. It achieved an F1 score of 0.87 and a precision of 0.90 on fraudulent transactions, proving its ability to detect fraud. However, there are 5,364 transactions misclassified by Logistic Regression, which did not maximize the simplicity and accuracy trade-off. Even with these errors, this model is less likely to overfit and may generalize better to new data. Therefore, the Random Forest Classifier is found to be the best model with perfect performance on all metrics.

## 4.4 Summary

The results chapter provides the evaluation of “machine learning models” used to detect fraud, namely Random Forest, Decision Tree, XG Boost and Logistic Regression. Logistic Regression got 95% accuracy while the random forest model got 100%. Hence, the best model has been identified as the “random forest classifier” model.

# References

Alsuwailem, A.A.S., Salem, E. and Saudagar, A.K.J., 2023. Performance of different machine learning algorithms in detecting financial fraud. *Computational Economics*, *62*(4), pp.1631-1667.

Alturki, R., 2021. Research Onion for Smart IoT‐Enabled Mobile Applications. Scientific programming, 2021(1), p.4270998.

Bodepudi, H., 2021. Credit card fraud detection using unsupervised machine learning algorithms. Int J Comput Trends Technol, 69, pp.1-13.

Hensen, B., Mackworth-Young, C.R.S., Simwinga, M., Abdelmagid, N., Banda, J., Mavodza, C., Doyle, A.M., Bonell, C. and Weiss, H.A., 2021. Remote data collection for public health research in a COVID-19 era: ethical implications, challenges and opportunities. Health policy and planning, 36(3), pp.360-368.

Mehbodniya, A., Alam, I., Pande, S., Neware, R., Rane, K.P., Shabaz, M. and Madhavan, M.V., 2021. [Retracted] Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques. *Security and Communication Networks*, *2021*(1), p.9293877.

Njoku, D.O., Iwuchukwu, V.C., Jibiri, J.E., Ikwuazom, C.T., Ofoegbu, C.I. and Nwokoma, F.O., 2024. Machine learning approach for fraud detection system in financial institution: a web base application. *Machine Learning*, *20*(4), pp.01-12.

Njoku, D.O., Iwuchukwu, V.C., Jibiri, J.E., Ikwuazom, C.T., Ofoegbu, C.I. and Nwokoma, F.O., 2024. Machine learning approach for fraud detection system in financial institution: a web base application. *Machine Learning*, *20*(4), pp.01-12.

Pallavi, C., Girija, R., Vedhapriyavadhana, R., Dey, B. and Vincent, R., 2021. A relative investigation of various algorithms for online financial fraud detection techniques. In Recent Trends in Intensive Computing (pp. 22-32). IOS Press.

Purohit, N. and Vishwakarma, R.G., 2021. Credit Card Fraud Detection Using Machine Learning Algorithms Using Python Technology. *Webology (ISSN: 1735-188X)*, *18*(6).

Vuppula, K., 2021. An advanced machine learning algorithm for fraud financial transaction detection. *Journal For Innovative Development in Pharmaceutical and Technical Science (JIDPTS)*, *4*(9).