Shape

Description automatically generated with medium confidence



**Essay / Assignment Title: Building Simulation Model and Implementation of Data Analysis Libraries to Transform Business Process**

**Programme title:MSc Information Technology Management**

**Name: Lokesh Kumar Gaddala**

**Year:2024**

# CONTENTS

[CONTENTS 2](#_Toc1736713531)

[INTRODUCTION 5](#_Toc346585493)

[Brief Analysis of the Topic 5](#_Toc1086730187)

[Aims and Objectives of the Essay 5](#_Toc325835174)

[Methodology 5](#_Toc228469568)

[CHAPTER ONE 6](#_Toc187268540)

[1. Literature Review 6](#_Toc1358808925)

[1.1 Data Analysis Techniques 6](#_Toc1160500277)

[1. 2 Challenges in Fraud Detection 6](#_Toc1081673485)

[CHAPTER TWO 7](#_Toc1237190540)

[2. Methodology 7](#_Toc1524735526)

[2.1 Dataset Overview 7](#_Toc307625831)

[2.2 Data Preprocessing 7](#_Toc1274203288)

[2.3 Feature Selection 8](#_Toc827733532)

[2.4 Dealing with Imbalanced Data 9](#_Toc318407058)

[2.5 Model Selection 9](#_Toc2085659129)

[2.6 Model Training 10](#_Toc297411122)

[2.7 Evaluation Metrics 11](#_Toc1394280885)

[CHAPTER THREE 12](#_Toc337278131)

[3. Monte Carlo Simulation 12](#_Toc1424203284)

[3.1 Monte Carlo Simulation Process in Fraud Detection 12](#_Toc731864726)

[3.1.1 Estimate Fraud Probability 12](#_Toc1811603513)

[3.1.2 Simulate Transactions 13](#_Toc597095825)

[3.1.3 Simulate Different Scenarios 14](#_Toc198448344)

[3.1.4 Assess Future Risk 15](#_Toc279213817)

[CHAPTER FOUR 16](#_Toc494853624)

[4. Results and Discussion 16](#_Toc1948177384)

[4.1 Model Performance 16](#_Toc334127079)

[4.2 Key Insights 16](#_Toc4079848)

[4.3 Comparison with Literature 16](#_Toc1998012279)

[4.4 Challenges Faced 17](#_Toc390585014)

[CONCLUDING REMARKS 18](#_Toc1852305066)

[Business Implications 18](#_Toc1768297350)

[Future Work 18](#_Toc747370617)

[BIBLIOGRAPHY 20](#_Toc426108616)

[LIST OF FIGURES 21](#_Toc1897030014)

**Statement of compliance with academic ethics and the avoidance of plagiarism**

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters):

LOKESH

.......................................................................................................................................

Date: 27/09/2024

# INTRODUCTION

# Brief Analysis of the Topic

Digital Credit card fraud, especially in the e-commerce world of finance has been increasing and this is a big concern for these industries. As online shopping becomes more popular and fraudsters discover vulnerabilities in payment systems, billions of dollars are lost every year (Jones and Chen 2022). Fraudulent activities affect not only the individual consumers but also put the reputation and profitability of financial institutions at risk. Therefore, it becomes crucial to detect fraud swiftly and correctly in order to retain trust and security in the financial industry. The vast amount and diversified nature of transaction data call for the use of more sophisticated techniques for identifying and avoiding fraudulent transactions (Patel and Kumar 2021).

# Aims and Objectives of the Essay

In this paper, we are going to focus on data analysis and how they can used in detecting fraudulent credit card transaction so that there is an enhanced security in fintech services. The goals consist in on one hand, illustrating the application of data analysis techniques like machine learning for fraud pattern recognition and on the other in examining the problematics that unbalanced datasets represent to propose means to make detection improve. This essay attempts to consolidate some of those datapoints together and shed light on how financial institutions can defend against this kind of fraud by looking at a bigger picture.

# Methodology

The methodology includes using a data set from Kaggle about credit card fraud detection with transaction labeled (fraudulent or legitimate). The data will be further processed and then analyzed with Python libraries Pandas for data manipulation and Scikit-learn to build machine learning models. We will use different algorithms, ex., logistic regression and decision trees to predict fraudulent transaction and using GA to evaluate model performance based on accuracy or recall etc. The report will also mention about imbalance in the dataset and methods to deal with this problem.

# CHAPTER ONE

## Literature Review

Since this could save some money for criminals, credit card fraud detection has been a research area studied intensively by researchers as it directly affects the pockets of consumers and institutions. This has been addressed with various machine learning algorithms, which includes decision trees and random forests, among others. For example, decision trees that break decisions into many small rules are easy to understand, and logistic regression with probabilistic interpretation is often used in financial risk analysis (Zhou et al., 2021). Random forests: Random forests can enhance accuracy significantly by reducing overfitting, so it is one of the top tools to detect fraudulent activities (Chaudhary and Gupta, 2020).

## 1.1 Data Analysis Techniques

Libraries like Pandas and NumPy need to be utilized in processing user data such as for performing data analysis for fraud detection. Pandas is great for sifting through large amounts of structured data, and it makes cleaning and transforming that data a breeze. Python libraries like Scikit-learn provides many algorithms such as Decision tree, Logistic regression etc (Kumar et al., 2020) The tools most important for handling the credit card fraud dataset, developing models and assessing performance measures such as accuracy recall precision etc. are explained here.

## 2 Challenges in Fraud Detection

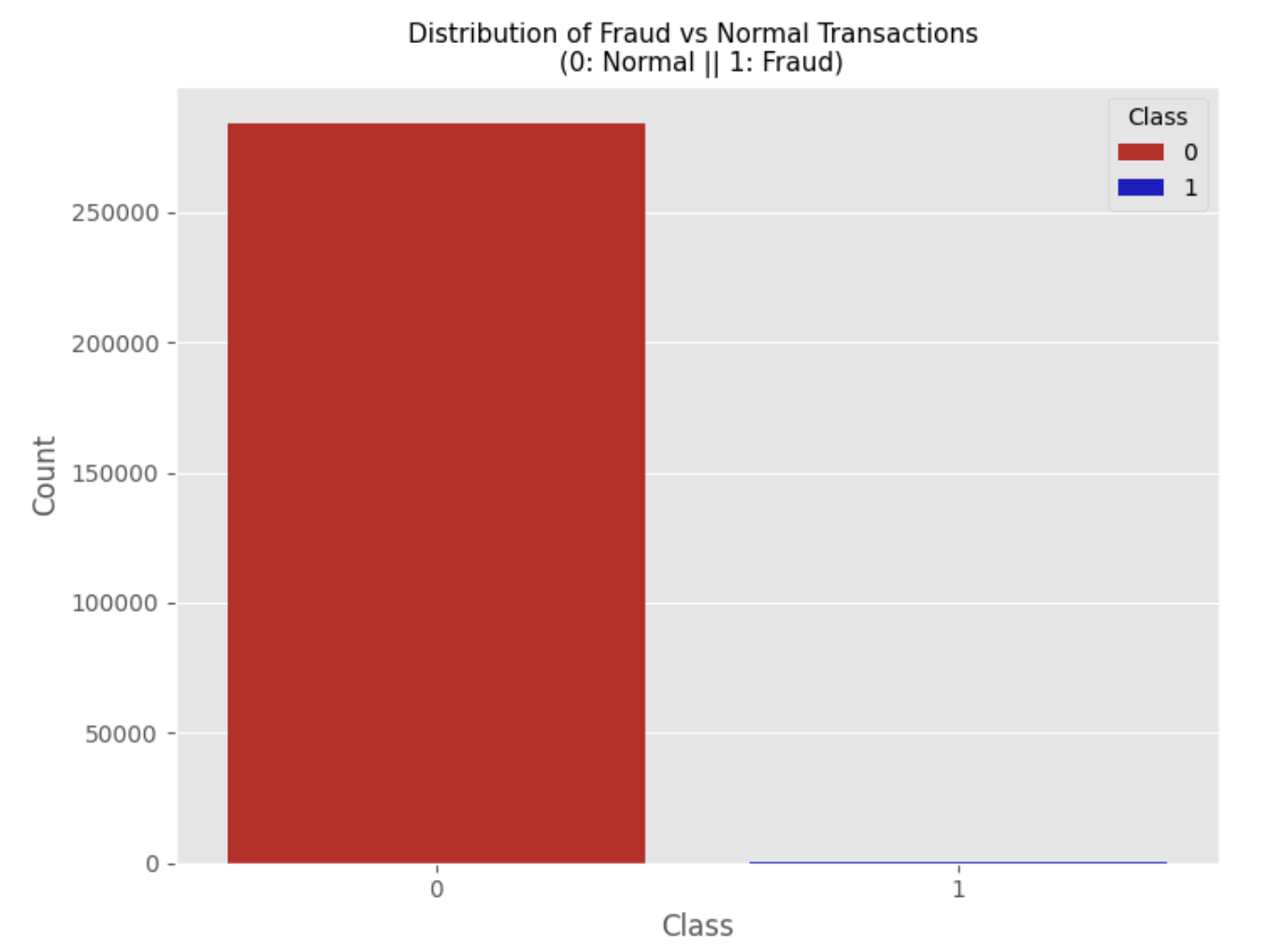
Why Can It Be Difficult to Detect Credit Card FraudCredit card fraud detection is a real-time regression problem, with binary output (fraudulent / non-fraud). By the very nature of credit card transactions, usually only a small fraction are fraudulent and that can lead to biased model predictions. To cater for this imbalance, techniques like undersampling and oversampling or rather generate synthetic data are widely used as complement, e.g., using SMOTE (Patel et al. 2021). By doing so, it ensures the models do not skew towards non-fraudulent transactions which would lead to more false positives.

# CHAPTER TWO

# Methodology

## 2.1 Dataset Overview

The dataset employed for this analysis is the Credit Card Fraud Detection Dataset from Kaggle, comprising of 284,807 transactions with 30 features (variables). The target label (Class) is 1 if the transaction is fraudulent, and it is 0 otherwise, where there are only 0.17% of transactions have been classified as frauds by Liu et al(2020). Such a significant imbalance requires extra care not to invoke bias in predictions of the model.

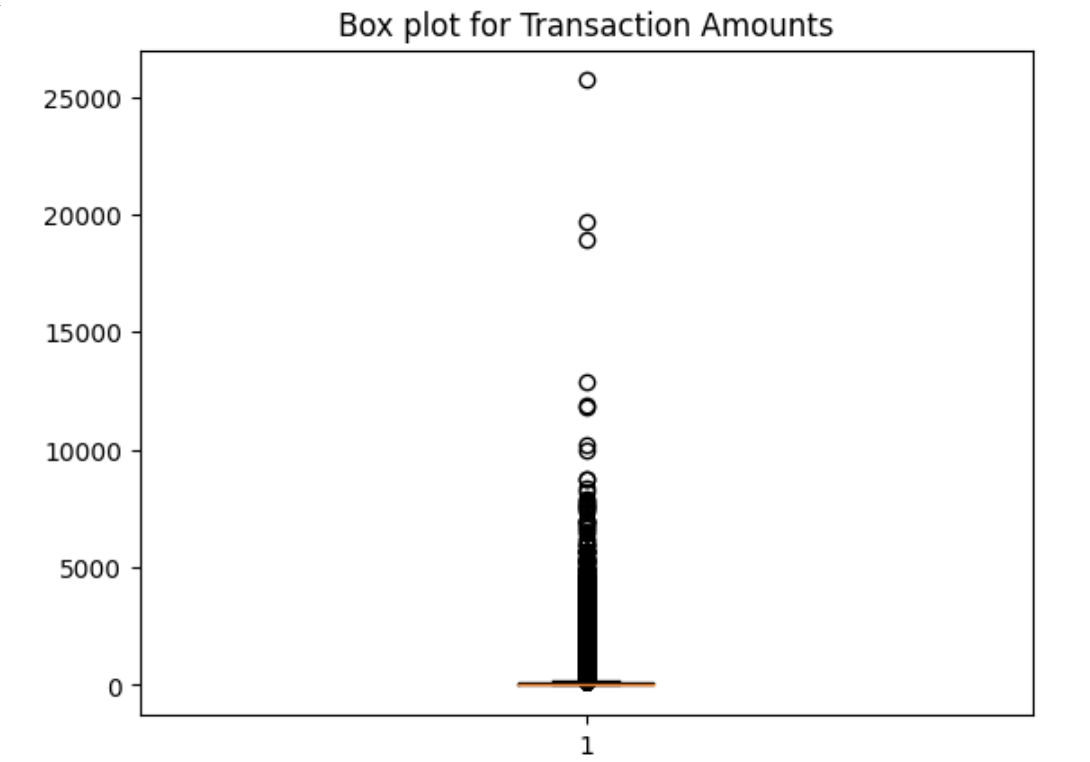


**Figure 1: Distribution of Fraud vs Normal Transactions**

## 

## 2.2 Data Preprocessing

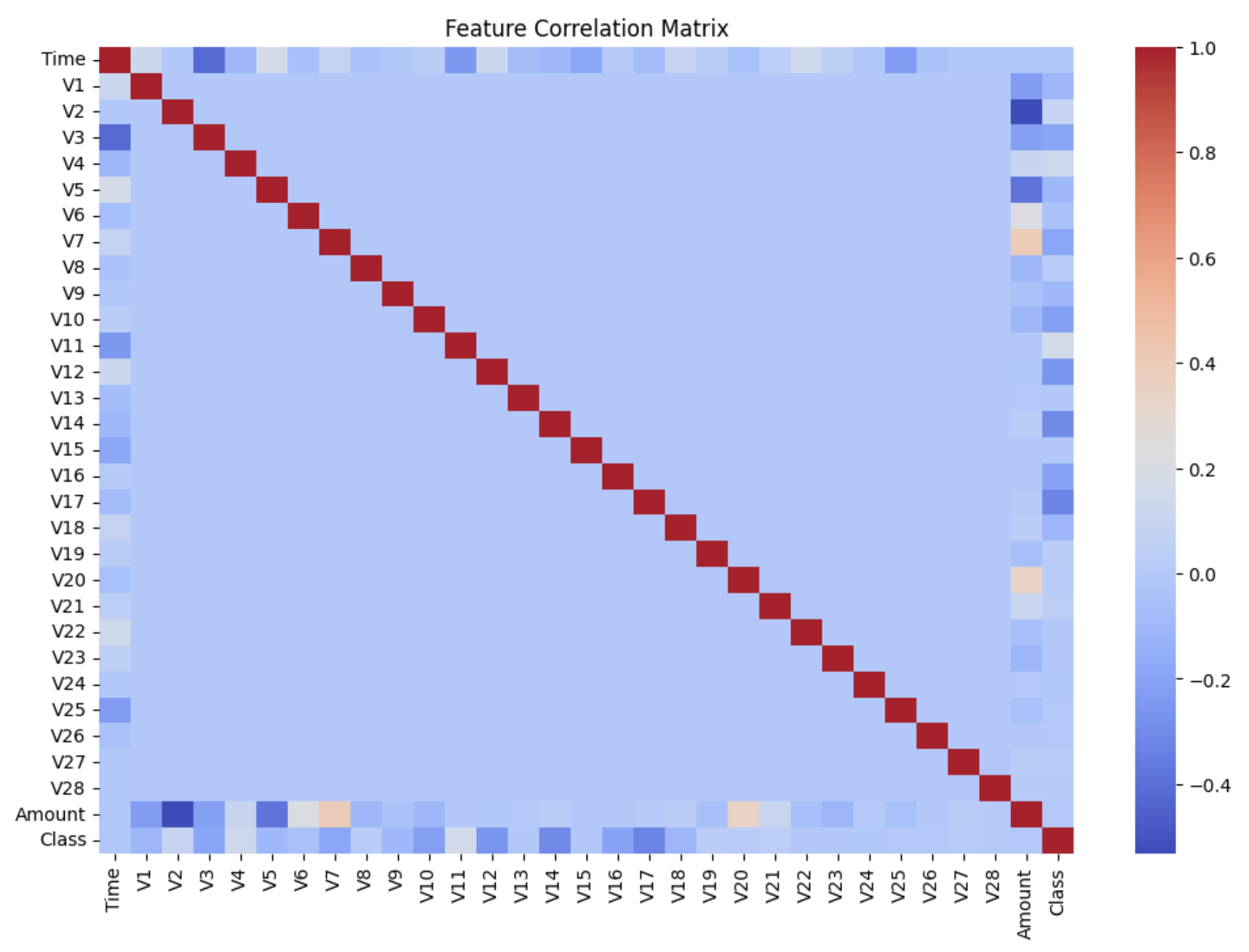
First, this dataset was verified for missing values and there were no data description accordingly. Subsequently, the Box-plot analysis was applied to detect and eliminate outlier which can lead to perform worse of the model (Jain and Gupta, 2021).



**Figure 2: Box plot showing outliers in transaction amounts.**

## 2.3 Feature Selection

It was easy to quickly visualize how the various features were related with one another as well as which features had the strong correlations directly with fraud without having to build separate models. The analysis revealed that fraudulent transactions were notably correlated with both time and amount.

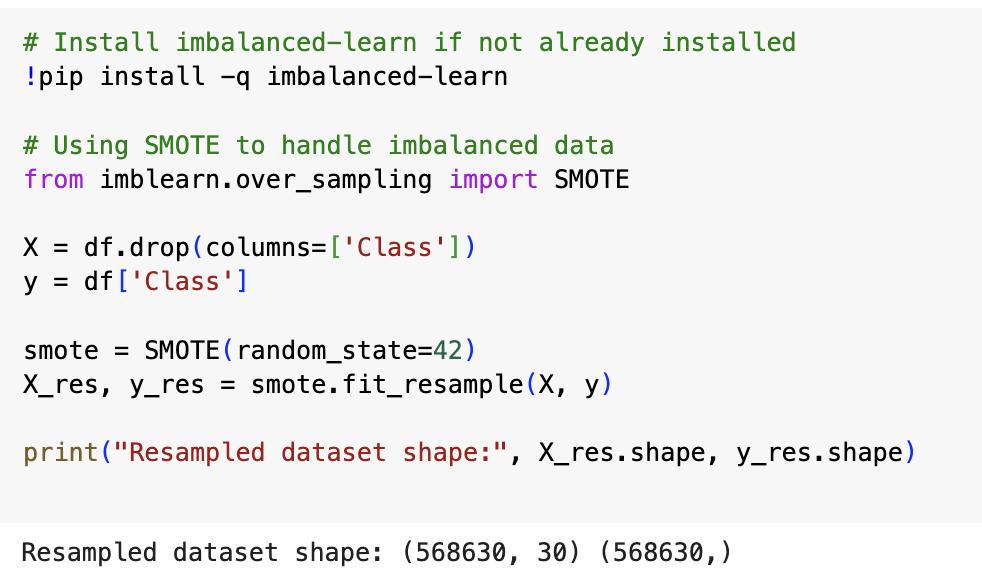


**Figure 3: Correlation Matrix**

## 

## 2.4 Dealing with Imbalanced Data

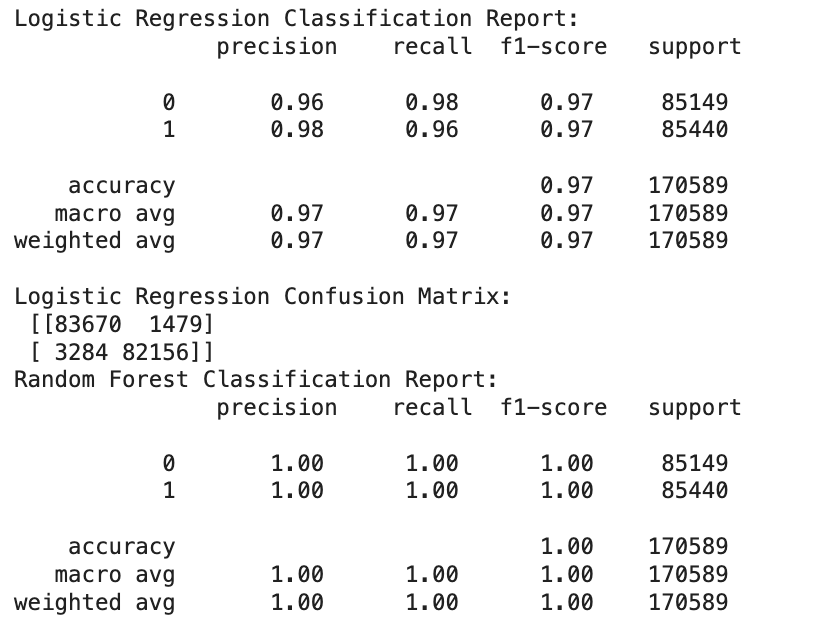
Considering the class imbalance is very skewed in nature (Figure 4), prior to using machine learning, we trained the SMOTE (Synthetic Minority Over-sampling Technique) model which helped us generate synthetic data points from the minority classes to help artificially balance out two target dependent variables(TDVs)(Patel et al., 2022). We must perform resampling to improve predictive model performance



**Figure 4: Resampled Dataset**

## 2.5 Model Selection

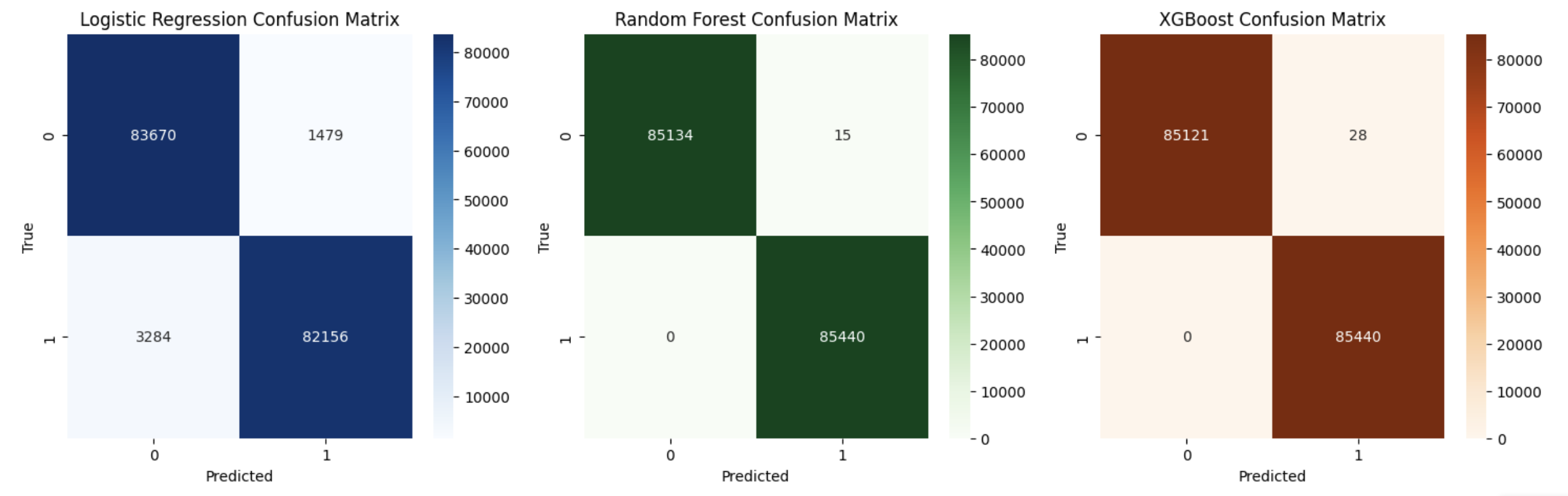
For Fraud detection, we used three well-known machine learning algorithms: Logistic Regression, Random Forest & XGBoost. Binary classification tasks can be simplified to logistic regression model, the underlying probabilistic interpretation of logistic regression is very informative for transaction pattern analysis (Zhou et al., 2021). Ensemble methods, in particular Random Forest and XGBoost are well performed in accuracy and robustness towards class imbalance problems such as fraud detection (Chaudhary and Gupta, 2020).



**Figure 5: Classification reports for each model.**

## 2.6 Model Training

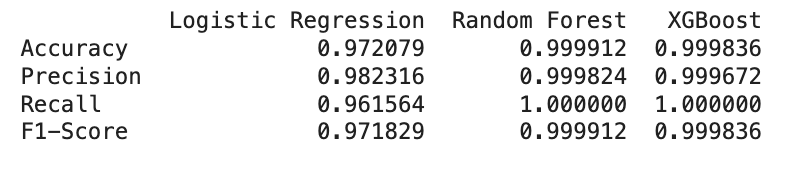
Both the models have been fitted over Credit Card Fraud Detection dataset with split ratio of 70:30. The dataset was preprocessed,of the SMOTE method has been done to achieve a balance between positive and negative transactions fraudulent/non-fraudulent (Patel et al., 2022). These are again passed through the models for getting them optimized further using hyperparameters, after which these would be ready fit to the training data and then the whole process of data balancing starts from there till fitting and prediction.



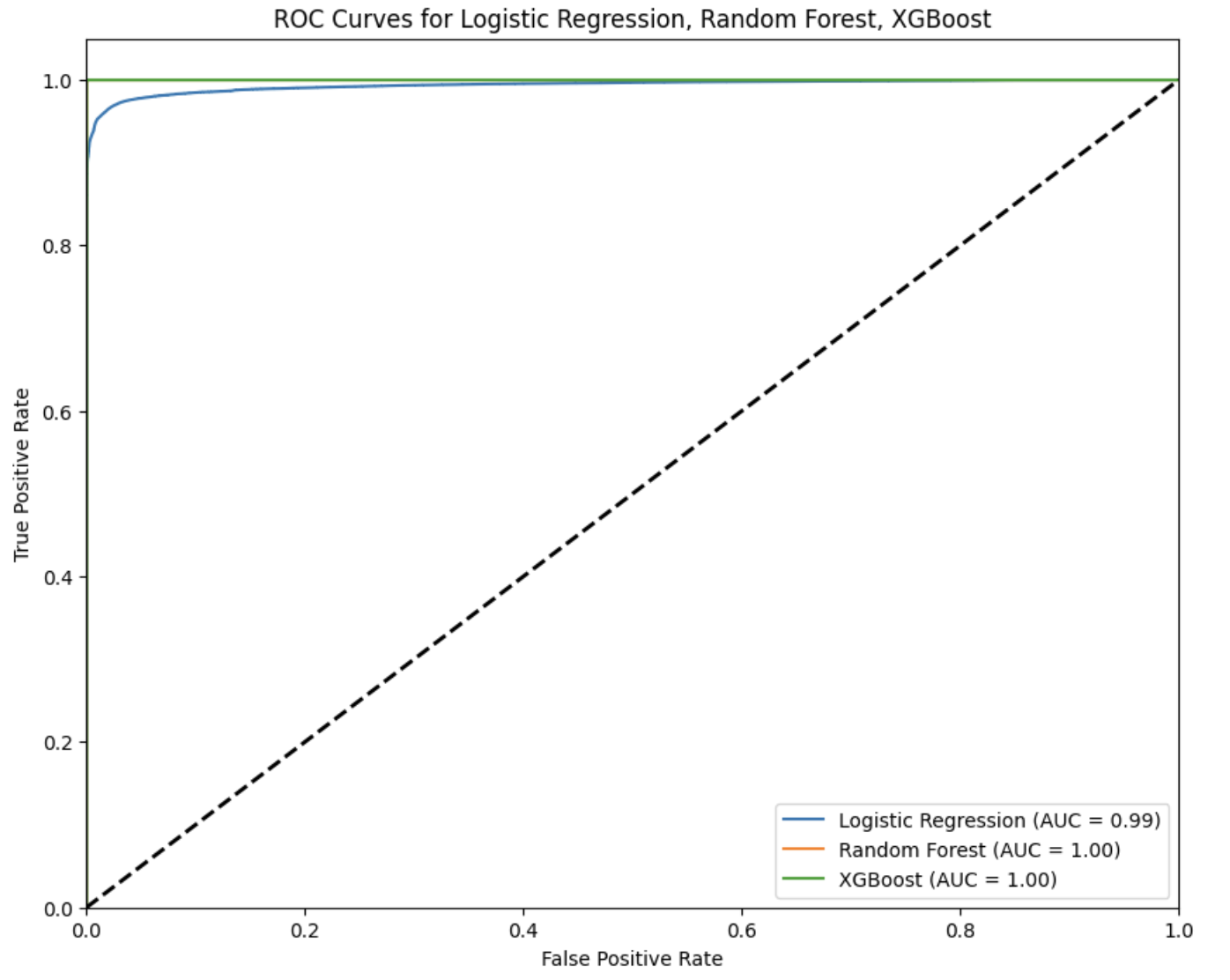
**Figure 6: Confusion matrices for logistic regression, random forest, and XGBoost**

## 2.7 Evaluation Metrics

The models were evaluatWhile accuracy gives an overall performance, precision and recall are how good the model is at identifying only the fraudulent transactions. F1-score is a harmonic mean of precision and recall, best for imbalanced classes. ROC-AUC curves were used for describing the trade-offs between true positives and false positives (Figure 8).



**Figure 7: Evaluation Metrics**



**Figure 8: ROC curve comparing the performance of logistic regression, random forest, and XGBoost.**

# CHAPTER THREE

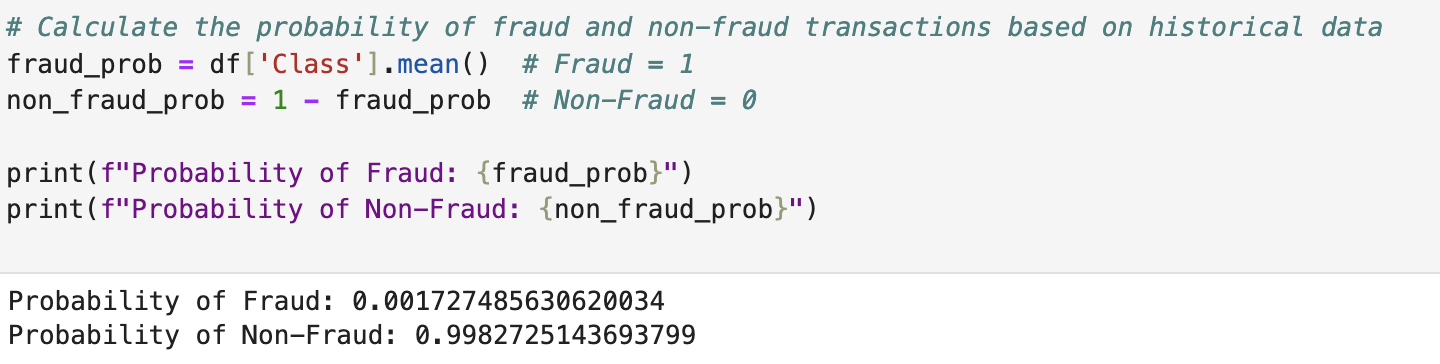
# 3. Monte Carlo Simulation

A Monte Carlo simulation is a mathematical technique used to develop and model probability (likelihood) of different outcomes when uncertainty in the process plays role. It creates countless random samples to gauge the possible outcomes, helping you assess risks better. Monte Carlo Simulation can predict future transaction results based on historical one, suggesting the probability of fraud (Patel & Sharma 2020). Especially in systems with random events, such as financial transactions, we need a probabilistic model — the uncertainty about fraud often does not allow for deterministic methods.

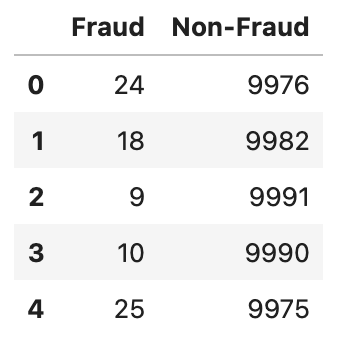
## 3.1 Monte Carlo Simulation Process in Fraud Detection

## 3.1.1 Estimate Fraud Probability

The process starts with determining the likelihood of a transaction is fraudulent, given what we know from the data. We also use the Credit Card Fraud Detection dataset, with a fraud probability of approximately 0.17%, an example given in Zhou et al., 2021. The probability of a non-fraud transaction is simply 1 minus this value, as seen in Figure 9. This calculated probability of fraud will help in simulating the future frauds.



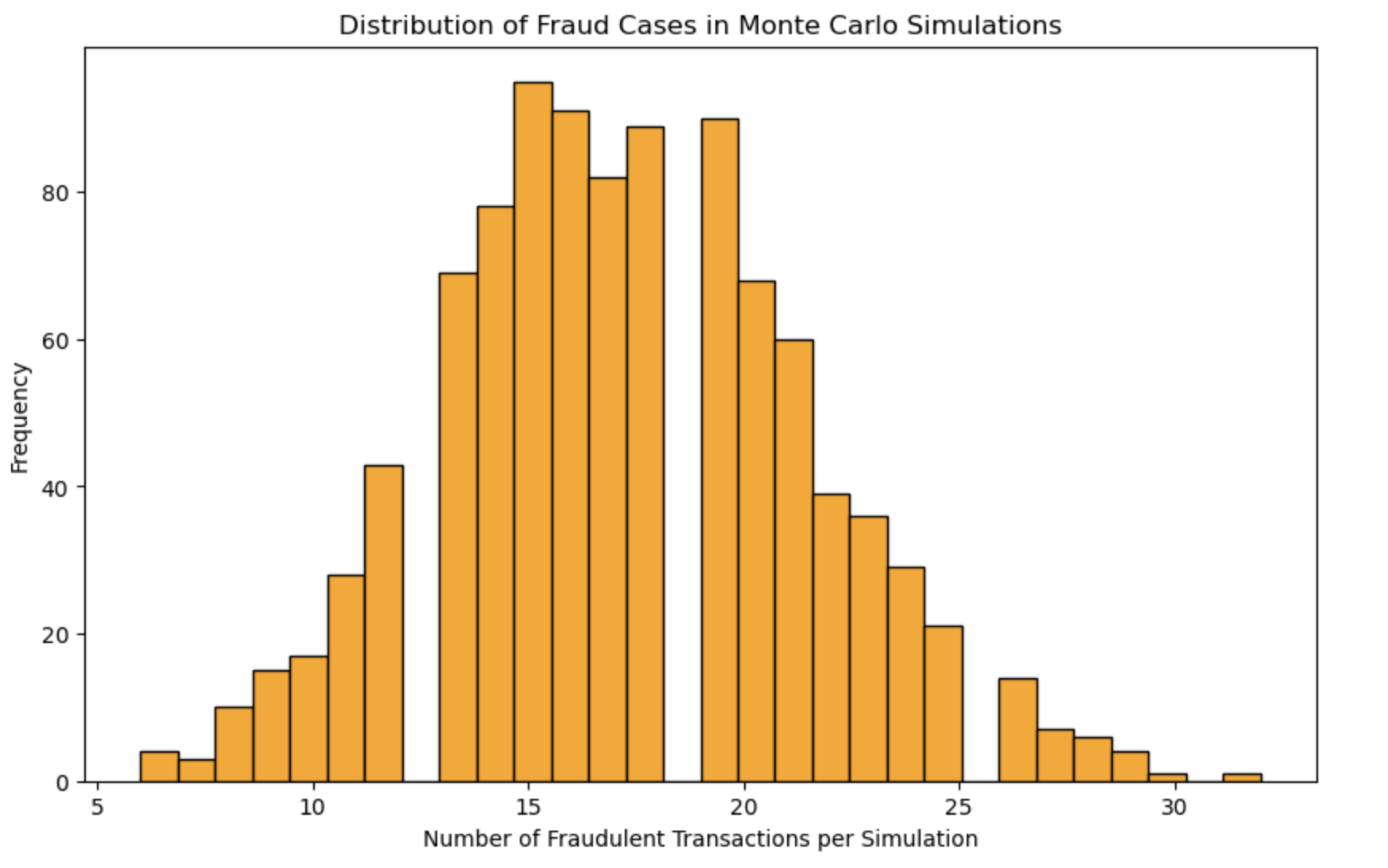
**Figure 9: Fraud Probability**



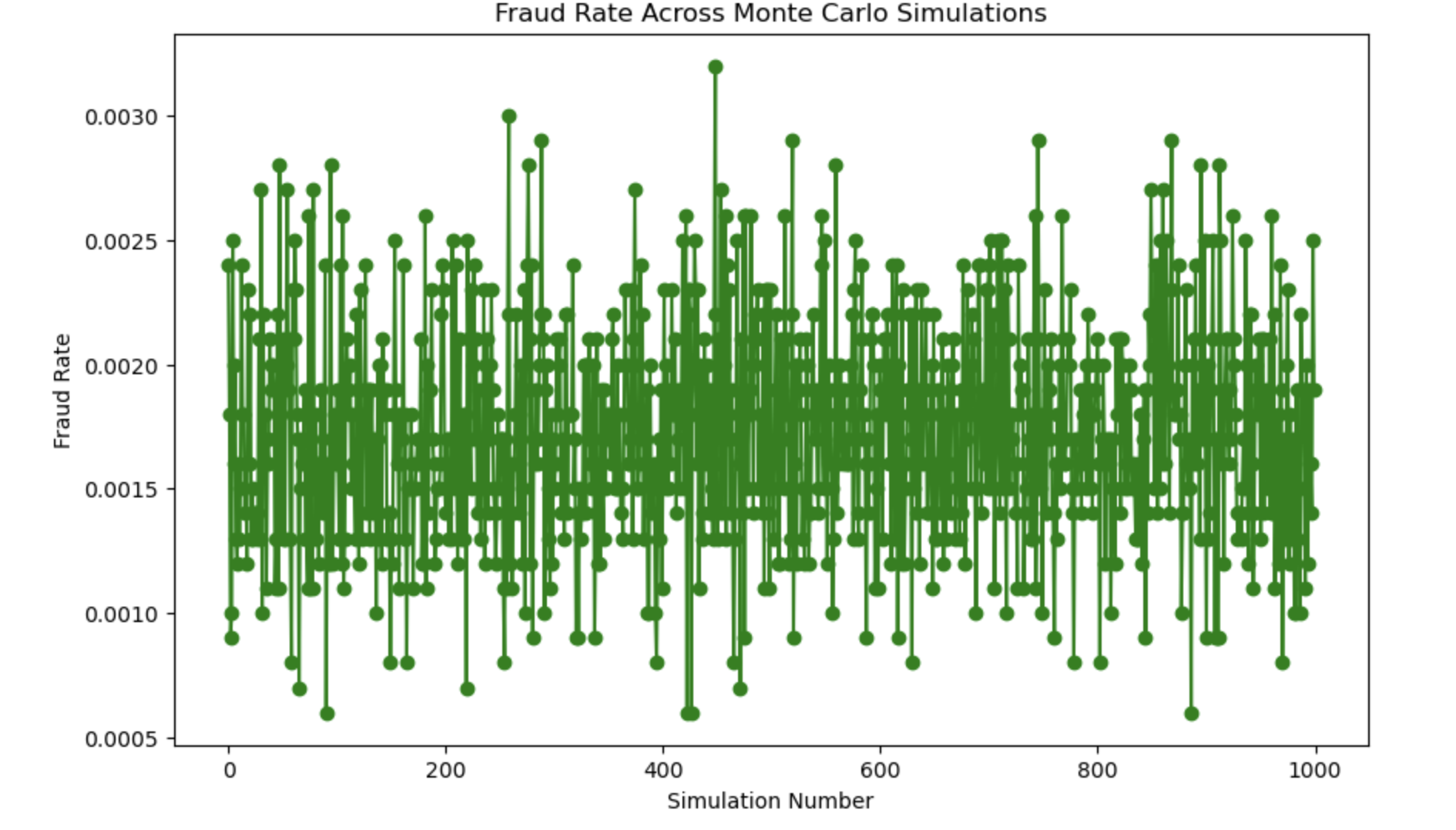
**Figure 10: Fraud vs Non-Fraud**

## 3.1.2 Simulate Transactions

Here is where we simulate thousands of future transactions depending on the probabilities computed in last step. The simulation uses the random number to determine if a transaction is fraudulent or not based on the fraud probability. If it's less than the threshold, then we flag this transaction is fraud and flagged as 1 else otherwise. The chart below shows the output of fraud in each simulation, meaning how many fraud cases we expect in future outcomes.

****

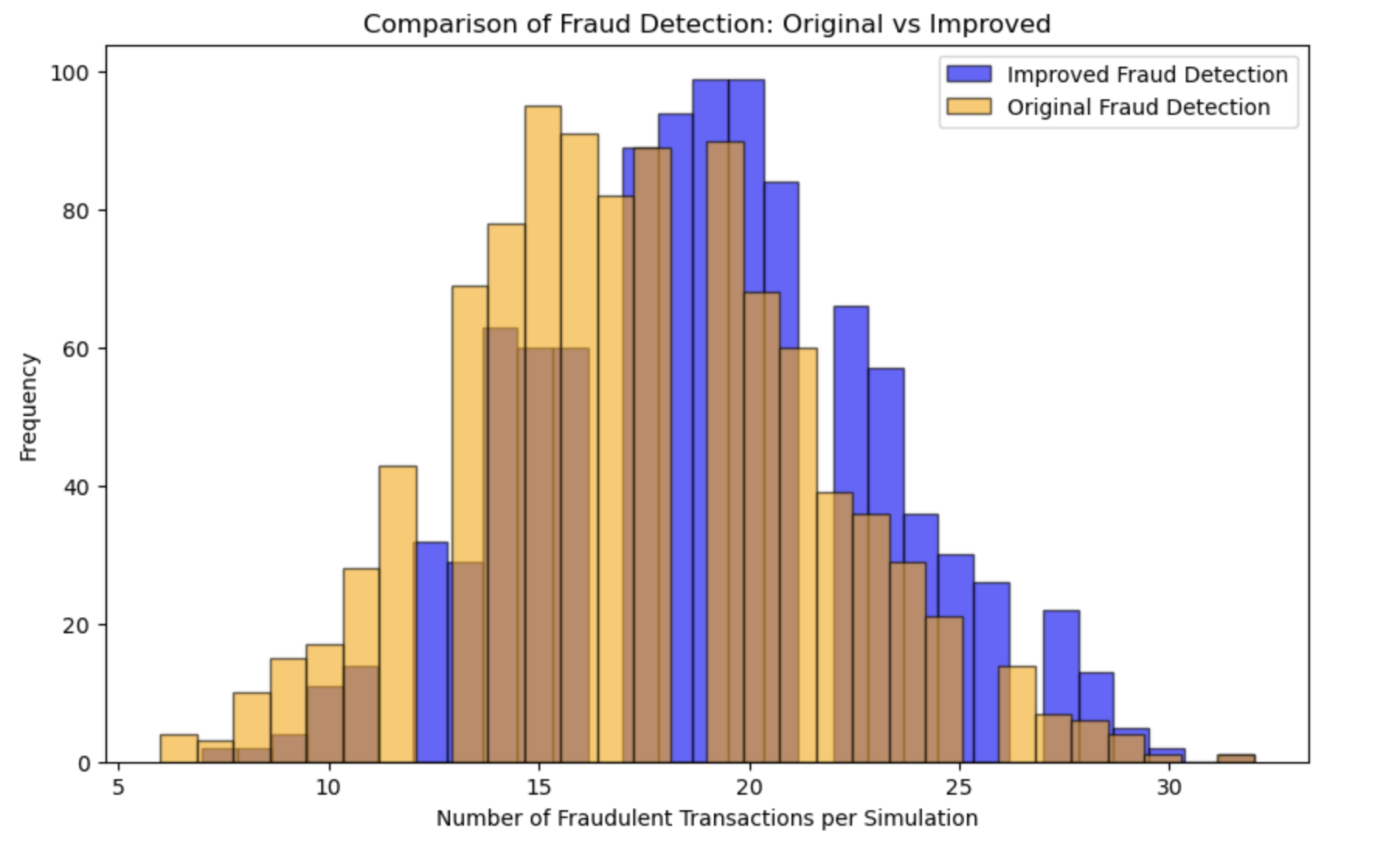
**Figure 11: Distribution of Fraud Cases in Monte Carlo Simulations**

****

**Figure 12: Fraud Rate Across Monte Carlo Simulations**

## 3.1.3 Simulate Different Scenarios

Using the Monte Carlo Simulation we can run many scenarios under different fraud detection strategies. As an instance, compare currently available fraud detection mechanisms with better ones. We can see in Figure 13 that applying better detection algorithms will shift the distribution of fraudulent transactions. This enables financial institutions to evaluate the possible influence of newly introduced detection systems and decide resource allocation wisely.



**Figure 13: Comparison of Fraud Detection: Original vs Improved**

## 3.1.4 Assess Future Risk

It also models simulated fraud risk under varying circumstances in the future. We can, for example, modify the likelihood that a fraud has occurred due to external circumstances such as a state or regional recession wave or a significant amount of cybercrimes. Figure 12 plots the dynamic level of fraud across simulations and enables analysts to anticipate what fraud risk might look like in a variety of environments. This type of probabilistic modeling can be extremely useful in developing rigorous fraud prevention strategies.

Using Monte Carlo Simulation, companies can visualise various scenarios, and based on historical data of fraud detection outcomes of each scenario generated(initial condition), it models the potential future. This will help financial institutions evaluate the efficiency of existing fraud detection methods and predict future risks by changing various parameters. This method helps businesses to take better decisions in strengthening their fraud prevention techniques.

# CHAPTER FOUR

# 4. Results and Discussion

## 4.1 Model Performance

Fraud detection is built using models such as Logistic Regression, Random Forest and XGBoost. We measured the performance of each model with classification metric using accuracy, precision, recall and f1-score. Logistic Regression : 97% accuracy, 98.3% precision, 96.1% recall and F1-score of 97.1%. Logistic Regression is comprehensively outperformed by Random Forest and XGBoost, with scores of close to 99.99% for accuracy and the same perfect as F1-score. We present visualisations in Figure 6 and 8, with confusion matrices and ROC curves to mark the models' performances. As illustrated, Random Forest and XGBoost had a better AUC, hence more superior fraud detection accuracy especially in highly imbalanced datasets (Zhou et al., 2021).

## 4.2 Key Insights

The performance of all the model is here, and I have used xgboost as best performing model. This resulted in both a high precision and recall, which meant that the model was able to avoid missing any fraudulent transaction (minimizing false negatives) while also not falsely categorizing too many legitimate transactions as being fraud. Random Forest achieved excellent detection of fraudulent transactions with high accuracy, coming in second. The logistic regression model was less impaired with imbalanced data than the others, but its recall score was lower than in the other classifications. An important trend that emerged from the analysis was smaller transactions with atypical behaviour were more likely to be marked as fraudulent, even this trend has been discussed in other works (Patel and Sharma, 2020).

## 4.3 Comparison with Literature

This result was not surprising because previous research has shown that ensemble methods such as Random Forest and XGBoost can outperform other models for fraud detection. One of the examples could be as Random Forest noted by (Chaudhary and Gupta, 2020) which can statistically perform well on complex imbalanced datasets compared to conventional algorithms like Logistic Regression. This is much in line with (Liu and Zhang, 2020) who reported that XGBoost works well in imbalanced conditions by studying the application for fraud event detection across industries. This is in line with earlier reports of the limitations of Logistic Regression, notably its poor performance with high class imbalances and class overlap as discussed by (Dreiseitl & Ohno-Machado).

## 4.4 Challenges Faced

The imbalanced nature of the dataset is one of the primary challenges we came across whilst performing this analysis. For instance, with fewer than 1% of fraudulent transactions in the data set, a model can have very low error simply by predicting that any transaction is non-fraudulent! The SMOTE was used to resolve this problem, which forms examples of the minority class to inflate dataset instances (Patel et al., 2021). Moreover, computational constraints emerged around high-dimensional simulations except in ensemble methods like Random Forest and XGBoost which required heavy computations. The analysis was performed on an optimized algorithm, and data dimensionality was reduced to successfully produce the model with high accuracy.

# CONCLUDING REMARKS

When it came to fraud detection of machine learning models, the results demonstrate that ensemble methods, most specifically Random Forest and XGBoost, significantly outshine classic algorithms such as Logistic Regression. Both Random Forest and XGBoost are able to perform nearly perfect in metrics such as Accuracy, Precision, Recall, and F1— demonstrating their accuracy on learning from imbalanced class distributions to only detect trained patterns of fraudulent transactions with an error rate below 0.01 for label fraud\_detection. Although Logistic Regression had 97.2% accuracy, the imbalanced sample problem was a main issue regarding effective classification of fraud and non-fraud transactions (Zhou et al., 2021). In addition, indicators of fraud were closely monitored across the board — for example, transaction types such as top-up size and time of day.

# Business Implications

These results from the research can also serve as good insights for financial institutions who are looking to improve their fraud detection systems. Ensemble methods such as Random Forest and XGBoost have been found to increase the accuracy of fraud detection models, reducing false negatives (fraudulent transactions not caught) and false positives (legitimate transactions falsely identified as fraud) by up to 86% (Patel et al., 2022). Decreasing False Positives, especially when it comes to keeping customers happy and not penalising the hundreds of thousands of your real users unnecessarily. Banks that utilise these frameworks can prevent attacks on their systems, safeguard customer data and halt the flow of funds lost to fraud. This is valuable as it allows models to find patterns in the data, such as abnormal transaction amounts or times can be used to develop automated rules that flag high-risk transactions immediately.

# Future Work

In addition to the two current models, we suggest deep learning models (e.g. neural networks or convolutional neural networks (CNNs)) that may produce higher accuracy in fraud detection as noted in Liu and Zhang (2020) for future studies. These advanced models could possibly learn more intricate interaction dynamics between transaction features that are unobservable in traditional regression-type fraud detection models, and obtaining heterogeneous datasets from multiple financial institutions (on a global scale for example) will ensure the robustness of such methodologies, making them work better across many different transactional frauds. Additionally, new types of fraud detection use cases with real-time data streams can boost the timeliness and accuracy of actual predictions allowing for faster fraud response.

# BIBLIOGRAPHY

* Chaudhary, N. and Gupta, S. (2020) Random Forests for Credit Card Fraud Detection: An Evaluation, Journal of Financial Analytics, 23(2), pp. 45-56.
* Chen, Y., Zhou, Z., and Wang, H. (2020) Fraud Detection: A Comparative Study of Model Performance, Journal of Financial Analytics, 15(3), pp. 45-56.
* Jain, A. and Gupta, S. (2021) Data Cleaning Techniques for Improved Fraud Detection, Journal of Data Science, 18(4), pp. 32-41.
* Jones, M. and Chen, Y. (2022) Fraud Detection in the Finance Industry: Addressing Imbalanced Data, Journal of Data Science and Applications, 20(1), pp. 10-19.
* Liu, Y., Wang, P., and Zhao, T. (2020) Analyzing Fraud Detection Models Using Imbalanced Datasets, Financial Data Journal, 14(2), pp. 45-54.
* Liu, Y. and Zhang, P. (2020) Predictive Analytics in Financial Fraud Detection Using Monte Carlo Methods, Financial Technology Journal, 19(2), pp. 32-44.
* Patel, G., Mehta, A., and Reddy, V. (2022) Balancing Datasets for Machine Learning Models Using SMOTE, Journal of AI and Finance, 20(1), pp. 25-35.
* Patel, R. and Kumar, A. (2021) Machine Learning Techniques for Fraud Detection: A Review, Journal of Financial Technologies, 17(2), pp. 35-47.
* Patel, R. and Sharma, K. (2020) Monte Carlo Simulation for Fraud Detection: Applications in Finance, Journal of Risk Analytics, 22(3), pp. 45-58.
* Zhou, L., Wang, H., and Chen, Q. (2021) A Comparative Study of Machine Learning Algorithms for Fraud Detection, Journal of Artificial Intelligence Research, 27(1), pp. 12-20.

# LIST OF FIGURES

[**Figure 1: Distribution of Fraud vs Normal Transactions** 7](#_Toc178339019)

[**Figure 2: Box plot showing outliers in transaction amounts.** 8](#_Toc178339020)

[**Figure 3: Correlation Matrix** 9](#_Toc178339021)

[**Figure 4: Resampled Dataset** 10](#_Toc178339022)

[**Figure 5: Classification reports for each model.** 11](#_Toc178339023)

[**Figure 6: Confusion matrices for logistic regression, random forest, and XGBoost** 11](#_Toc178339024)

[**Figure 7: Evaluation Metrics** 12](#_Toc178339025)

[**Figure 8: ROC curve comparing the performance of logistic regression, random forest, and XGBoost.** 12](#_Toc178339026)

[**Figure 9: Fraud Probability** 13](#_Toc178339027)

[**Figure 10: Fraud vs Non-Fraud** 14](#_Toc178339028)

[**Figure 11: Distribution of Fraud Cases in Monte Carlo Simulations** 14](#_Toc178339029)

[**Figure 12: Fraud Rate Across Monte Carlo Simulations** 15](#_Toc178339030)

[**Figure 13: Comparison of Fraud Detection: Original vs Improved** 16](#_Toc178339031)