Identifying Fraudulent Transactions using Machine Learning: A Study of Credit Card Fraud

MOON KARMAKAR

ID- 40389123

Word Count- 8024

Research Report submitted in part
fulfilment of the degree of Master of Science in
Business Analytics

Year of Submission: 2023

Queen’s Management School

DECLARATION

This is to certify that:

i. The portfolio comprises only my original work;

ii. AI technologies (e.g. chat GTP) have not been used in the writing of the portfolio dissertation.

iii. No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Moon Karmakar.

[Candidate’s Signature]

Moon Karmakar

Printed Name

07 September 2023

Date

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who contributed to the successful completion of this project. First and foremost, I extend my deepest appreciation to my mentor Dr. Byron Graham and Dr Qiao (Olivia) Peng, whose guidance, support, and expertise were invaluable throughout this endeavor.

I would also like to thank my fellow classmates for their insightful discussions, which provided diverse perspectives and enriched the project.

Furthermore, I am grateful to the library staff and online resources for providing access to the wealth of information needed for my research.

Lastly, I want to acknowledge my family for their unwavering encouragement and understanding during this academic journey.

ABSTRACT

Credit card fraud is a type of financial misconduct in which unauthorized parties use credit card information to acquire products and services without the cardholder's permission. In recent years, the rise in online transactions has led to an increase in fraudulent activities. Annually, these fraudulent activities result in significant financial losses for businesses. Given that only a small proportion of credit card transactions are classified as fraudulent, the datasets used for fraud detection are inherently unbalanced, making detection difficult. This study focuses on the efficacy of ensemble methods that rely on tree structures for detecting fraudulent transactions. In this investigation, numerous classifiers, namely Random Forest, Bagging, XGBoost, LightGBM, Adaboost, and CatBoost, are employed. Combining random undersampling (RUS) and Borderline-SMOTE techniques, a combined sampling strategy is utilized to address the problem of class imbalance. In terms of fraud detection, the results indicate that boosting classifiers outperform bagging classifiers. Specifically, when metrics such as F1-Score, Matthews Correlation Coefficient (MCC), and Area Under the Precision-Recall Curve (AUC-PR) are considered, XGBoost exhibits significantly superior performance.

Keywords- Machine Learning, Credit Card Fraud, Class Imbalance, Borderline-SMOTE, Deep Learning Adaboost, LightGBM, Random Undersampling (RUS), and XGBoost.

Table of Contents

Tables

Figures

1. Introduction

1.1 Background

In the modern digital era, credit cards have become the predominant payment method for online purchases. In recent years, both the utilization of credit cards and the incidence of fraudulent transactions have increased dramatically. Credit card fraud is a significant and escalating problem that affects not only the financial industry but also the global economy. Fraud is a multimillion-dollar industry that grows annually. Globally, fraud imposes significant costs on our economy. For detecting fraudulent credit card transactions, modern techniques based on Data mining, Machine learning, Sequence Alignment, Deep Learning, Genetic Programming, Artificial Intelligence, etc. have been implemented(Chaudhary et al., n.d.). According to the most recent data from the Nilson Report, global credit card fraud losses reached $28.58 billion in 2020, a significant increase from the $23.97 billion reported in 2017(*Nilson Report - Google Search*, n.d.). These numbers highlight the importance of promptly identifying fraudulent transactions.

The prevention of credit card fraud is a top priority for financial institutions. While both supervised and unsupervised machine learning techniques can help identify credit card fraud, this study focuses solely on supervised classification techniques. In supervised fraud detection, transactions are classified as legitimate or fraudulent based on an analysis of historical data. Although fraudulent transactions are uncommon, their omission can result in significant financial liabilities. In addition, misclassifying a legitimate transaction as fraudulent incurs significant costs, tarnishing a company's reputation and alienating its most devoted customers. This emphasizes the need for a sophisticated credit card fraud detection system that precisely flags fraudulent transactions. By employing classification algorithms that identify patterns and deviations in historical data, the likelihood of identifying a fraudulent transaction prior to its actualization is increased. Figure 1 depicts a supervised framework for detecting credit card fraud that integrates data mining and machine learning.

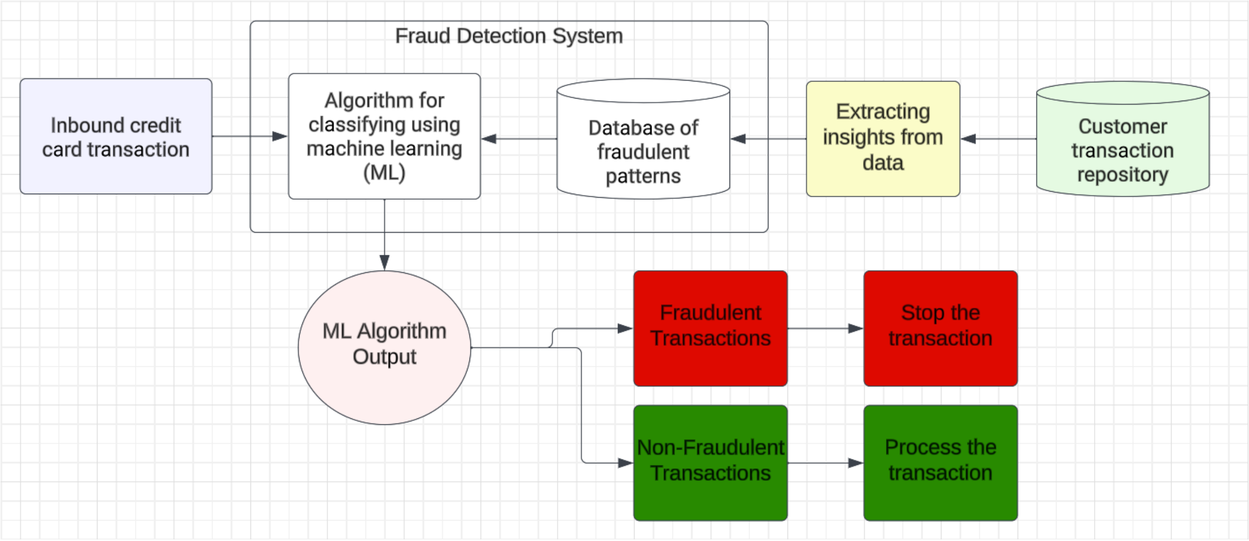

Fraud Detection System
Inbound credit card transaction
Algorithm for classifying using machine learning (ML)
Database of fraudulent patterns
Extracting insights from data
Customer transaction repository
ML Algorithm Output
Fraudulent Transactions
Stop the transaction
Non-Fraudulent Transactions
Process the transaction

Figure 1: Fraud Detection Infrastructure

The motive of this research is to develop a predictive model that can assist financial institutions in identifying potentially fraudulent transactions. A significant challenge in fraud detection is striking a compromise between minimizing the incorrect identification of legitimate transactions as fraudulent (false positives) and failing to detect actual fraudulent transactions (false negatives). Given that most credit card transactions are legitimate and only a small percentage involve fraud, credit card fraud datasets display a significant imbalance. Consequently, the task of developing dependable credit card fraud detection models becomes more complex. Another obstacle in this field is the constant evolution of fraudsters' strategies and changes in consumer behavior, also known as "concept drift." As those intent on committing fraud continually seek out new and innovative methods, the implementation of advanced technologies is required to remain ahead of these altering fraud trends and alterations in consumer purchasing habits.

A significant limitation encountered in fraud detection research is the dearth of publicly accessible authentic data in the domain of credit card fraud. Due to the unavailability of authentic datasets relating to credit card fraud, researchers frequently rely on synthetic data.

1.2 Scope of Study, Research Questions and Aims

The aim of this study is to examine the efficacy of tree-based ensemble algorithms in fraud prediction. This study's primary question is as follows:

Research Questions:

RQ1: Which algorithm is most accurate in predicting fraudulent transactions?

RQ2: Which factors are most important in predicting fraudulent transactions?

RQ3: What level of accuracy do tree-based ensemble learning algorithms demonstrate in distinguishing between legitimate and fraudulent credit card transactions by analysing past transactional data?

Aims:

Aim 1: Evaluate the performance of different machine learning algorithms in detecting credit card fraud.

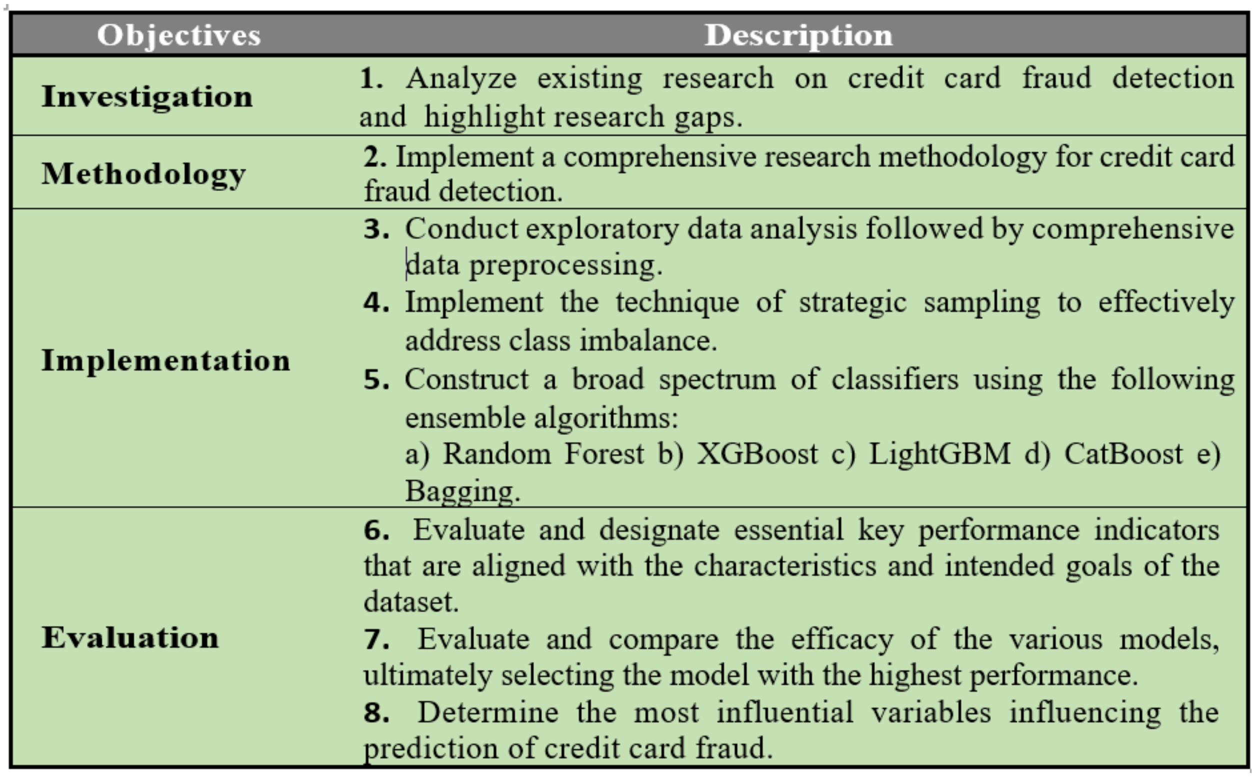
Aim 2: To investigate the impact of feature selection and dimensionality reduction techniques on the performance of fraud detection models.

Aim 3: To investigate the ethical considerations and privacy implications associated with credit card fraud detection using machine learning.

Aim 4: To develop a comprehensive framework for real-time fraud detection in credit card transactions.

To comprehensively investigate the research inquiry, a distinct set of research aims has been outlined and presented in Table 1.

Table 1: Objectives and Description



Objectives	Description
Investigation	**1.** Analyze existing research on credit card fraud detection and highlight research gaps.
Methodology	**2.** Implement a comprehensive research methodology for credit card fraud detection.
Implementation	**3.** Conduct exploratory data analysis followed by comprehensive data preprocessing. **4.** Implement the technique of strategic sampling to effectively address class imbalance. **5.** Construct a broad spectrum of classifiers using the following ensemble algorithms: a) Random Forest b) XGBoost c) LightGBM d) CatBoost e) Bagging.
Evaluation	**6.** Evaluate and designate essential key performance indicators that are aligned with the characteristics and intended goals of the dataset. **7.** Evaluate and compare the efficacy of the various models, ultimately selecting the model with the highest performance. **8.** Determine the most influential variables influencing the prediction of credit card fraud.

Because there is a scarcity of publicly available fraud datasets, a substantial portion of present research in fraud detection leans heavily on the European dataset, which predominantly comprises numerical input variables. The primary contribution of this research lies in its thorough assessment of the effectiveness of tree-based ensemble classifiers for fraud detection, facilitated by a substantial yet highly imbalanced dataset encompassing both continuous and categorical attributes. Furthermore, this study identifies the key predictors of credit card fraud.

2. Literature Review

Fraud is the intentional and illicit use of deception to obtain a financial or personal advantage. It entails deliberate actions that violate legal, regulatory, or policy frameworks in order to obtain unauthorized financial gains.

Extensive literature on the detection of anomalies or fraudulent activities in this field has been published in the past and is accessible to the public. A comprehensive study conducted by Clifton Phua and associates has revealed that this domain's strategies include data mining applications, automated fraud detection, and adversarial detection(Phua et al., 2010). Similarly, Suman, a Research Scholar at GJUS&T in Hisar HCE(Arafath et al., 2022), elaborated on credit card fraud detection techniques such as Supervised and Unsupervised Learning. While these methodologies and algorithms demonstrated unexpected successes in specific domains, they were unable to provide a uniform, long-lasting solution for fraud detection.

In a study by Nguyen et al., 2020, an analysis of diverse credit card datasets revealed that attribute quantity, transaction volume, and attribute correlations collectively influence the credit card fraud detection effectiveness of a model (Nguyen et al., 2020). Niu et al., 2019conducted a comprehensive investigation using the European dataset. Using the AUC-ROC metric, they conducted a comparative analysis of multiple models for detecting fraudulent transactions. The findings indicated that supervised methods were marginally superior to unsupervised ones. Notably, XGBoost and RF demonstrated the highest performance (AUC=0.989) among the supervised models(Niu et al., 2019). In contrast, Saito and Rehmsmeier (2015) argued that assessing binary classifiers on imbalanced datasets using the ROC curve could lead to erroneous conclusions due to its sensitivity to class imbalance. They supported the Precision-Recall (PR) curve, citing its superior informativeness and applicability as a performance metric for credit card fraud detection(Saito & Rehmsmeier, 2015).

Ensemble classifiers have demonstrated superior efficacy in credit card fraud detection compared to other contemporary classifiers. Using the European dataset, Dhankhad, Mohammed, and Far (2018) conducted a comparison of RF, Stacking classifiers (SC), LR, XGBoost, GBM, MLP, KNN, SVM, NB and Decision Tree (DT) (Dhankhad et al., 2018.). Notably, SC with LR as a meta-classifier, RF, and XGBoost produced the best results (F1-score = 0.95), leading to the conclusion that ensemble methods improve the effectiveness of fraud detection models. Husejinovic (2020) evaluated the efficacy of NB, Bagging and DT in fraud detection using the European dataset. The results demonstrated that Bagging with DT as the foundation learner had the maximum AUC-PR score of 0.825, followed by DT (0.745) and NB (0.080) with the lowest predictive capabilities (Husejinovic et al., 2020). Compared to Dhankhad et al. (2018), these results demonstrate the efficacy of tree-based ensemble techniques in detecting fraud. Zareapoor and Shamsolmoali (2015) examined the efficacy of NB, SVM, KNN, and bagged trees in fraud detection using an actual e-commerce transactions dataset (Zareapoor et al., 2015.). Based on their findings, Bagging with DT as the optimal classifier produced the highest Matthews Correlation Coefficient (MCC), the most robust fraud detection rate, and the lowest of rate false alarm. In addition, the ability of Bagging to manage class imbalance was highlighted (Zareapoor et al., 2015), thereby enhancing its efficacy. When compared to (Husejinovic et al., 2020)their findings demonstrate the optimistic potential of Bagging with DT as a weak learner in the context of credit card fraud detection.

Boosting is a prominent technique in the field of ensemble learning methods, alongside bagging. Divakar et al., 2019evaluated the European dataset with the aim of identifying the most effective boosting algorithm for credit card fraud prediction. The scope of their examination included AdaBoost, Gradient Boosting Machine (GBM), and XGBoost. XGBoost (reaching an F1-Score of 0.88) demonstrated superior performance compared to AdaBoost (F1-Score of 0.76) and GBM (F1-Score of 0.74) according to the findings of their study. These findings demonstrate the efficacy of boosting algorithms in detecting credit card fraud, corroborating the findings of Dhankhad et al., 2015, who similarly attested to the positive performance of XGBoost in detecting credit card fraud. Nevertheless, it is essential to recognize that the dynamic evolution of fraudulent patterns, which is characterized by concept drift, introduces the possibility of a progressive decline in the performance of cutting-edge fraud detection models over time.

In a study by Fang, Zhang, and Huang (2019), a comparative evaluation of the efficacy of LightGBM versus RF and GBM for fraud detection was conducted. While all models demonstrated commendable performance on a variety of datasets, LightGBM performed the best than RF and GBM in terms of AUC score and training time. This led to the conclusion that LightGBM is a particularly effective and efficient method for detecting fraud (Fang et al., 2019). In addition, Taha and Malebary (2020) investigated multiple datasets and introduced an optimized approach for fraud detection using LightGBM. This strategy entailed utilizing a Bayesian-based hyperparameter optimization algorithm to fine-tune the model's parameters and improve its performance. CatBoost is distinguished from conventional boosting algorithms by its unique capacity to manage categorical features autonomously(Taha & Malebary, 2020). (Hancock & Khoshgoftaar, 2020)evaluated the performance of CatBoost and XGBoost in the context of identifying fraudulent medical insurance claims. The research involved two datasets with a mixture of numerical and categorical attributes. The findings clearly demonstrated CatBoost's superiority over XGBoost, particularly when categorical features of substantial cardinality were involved.

Deep learning plays a crucial role in comprehending complex relationships between transactional attributes. Autoencoders (AEs) have a practical application in feature extraction for classification. (Misra et al., 2020)conducted an experiment using the European dataset and an AE to extract relevant characteristics from the input data. These characteristics were then fed into a classifier to determine the legitimacy of transactions. Using an AE followed by a Multi-Layer Perceptron (MLP) yielded the best F1-Score performance (0.8265), as determined by (Misra et al., 2020). Even though Artificial Neural Networks (ANNs) are more effective than conventional algorithms, their training process is slow. Jain, (2019) evaluated various supervised models utilizing the KDD'99 intrusion dataset. The findings demonstrated the superiority of ANNs over alternative models, but it was acknowledged that the computational requirements associated with training ANNs pose a significant challenge (Jain, 2019)

Existing research consistently demonstrates the effectiveness of ensemble learning techniques for augmenting the performance of credit card fraud detection models. It is evident, when contemplating performance evaluation, that the F1-Score is extensively used as a performance metric in this field. Chicco & Jurman, (2020), who examined the analysis of a profoundly unbalanced genomic dataset, offer an alternative viewpoint. For the evaluation of binary classifiers operating on highly skewed datasets, their study demonstrated that the Matthews Correlation Coefficient (MCC) produces results that are not only more accurate but also more informative than the F1-Score. This is a result of MCC's ability to consider all outcomes in the perplexity matrix as well as the total number of positive and negative observations.

3. Research Methodology

Given the objective of addressing a business challenge, a research approach with a scientific foundation, drawing partial inspiration from the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, has been formulated. Figure 2 visually illustrates this devised methodology aimed at addressing the research query introduced in Section 1.2. Subsequently, a comprehensive breakdown of the distinct phases is provided for clarity.

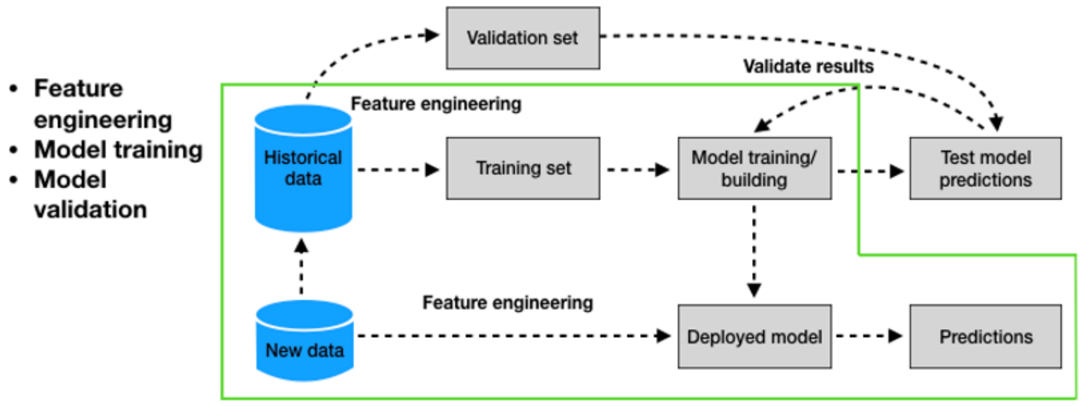


Figure 2: System Architecture for Credit Card Fraud Detection

Sourced:(*4. Baseline Fraud Detection System — Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook*, n.d.)

3.1 Business Problem Comprehension:

During this initial phase, a comprehensive understanding of the difficulty of credit card fraud detection is developed. This involves articulating the research query and project objectives. The objective is to illustrate the problem in the context of a data mining classification task, with the ultimate objective being the development of an ensemble classifier capable of identifying fraudulent transactions using historical transaction data.

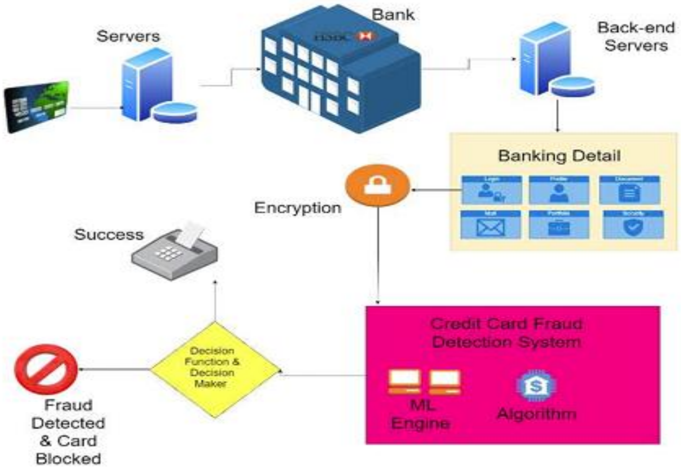


Figure 3: The Systematic Monitoring and Analysis of Financial Transactions

Sourced: (khan et al., 2021)

3.2 Exploratory Data Analysis (EDA)

An artificial dataset of credit card transactions has been utilized for the purpose of this study. During the period between January 1, 2019 and December 31, 2020, the dataset includes both legitimate and fraudulent transactions made by American cardholders. The dataset includes 1,852,394 credit card transactions from 1,000 consumers and 23 unique transaction attributes. Sourced from the Kaggle repository, this dataset serves as the foundation for predicting the authenticity of transactions. The outcomes of the binary classification problem are represented by the values '0' for 'legitimate' and '1' for 'fraudulent'. The original dataset was comprised of two distinct files, which were merged to provide a more comprehensive view of the data. This information is elaborated in Table 2, which provides insight into the characteristics of the dataset.

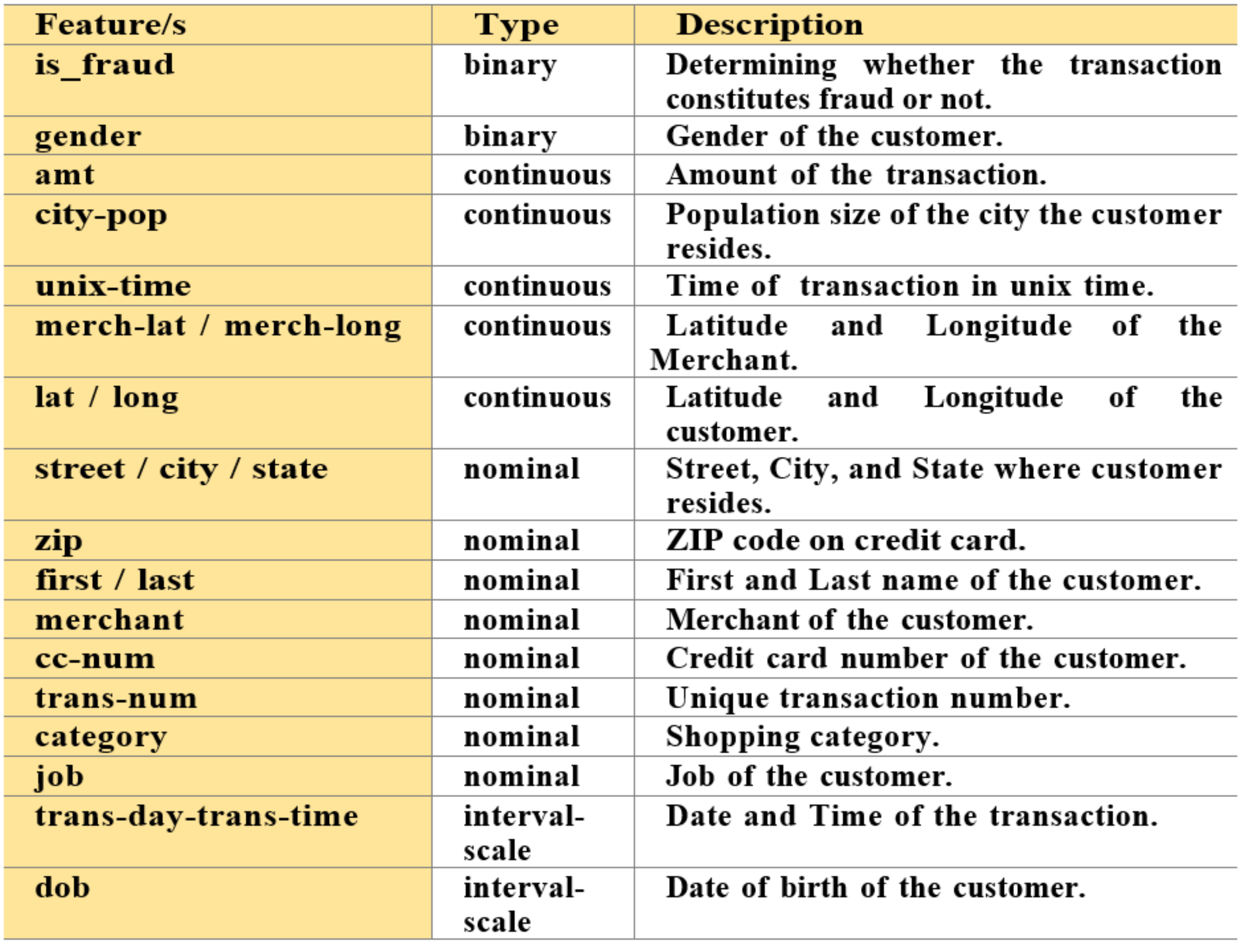
Exploratory Data Analysis (EDA) was performed to gain a deeper understanding of the inherent structure of the dataset. Figure 3 is a visual illustration of the volume and value disparity between legitimate and fraudulent transactions. Notably, the dataset is marked by a significant imbalance, with most transactions being legitimate and only a small percentage being fraudulent. Despite accounting for only 0.52 % of total transaction volume, fraudulent transactions account for 3.95 % of total transaction value.

In addition, a noteworthy observation reveals that the average transaction amount for illicit activities is $530.66, while the average transaction amount for legitimate transactions is substantially lower at $67.65. This disparity highlights the financial ramifications of failing to detect infrequent fraudulent transactions, which can result in significant losses. Particularly, the 'amount' variable for legitimate transactions contains outliers, whereas fraudulent transactions lack such anomalies. Notably, the dataset contains no missing values or duplicate entries, attesting to its reliability and validity.

3.3 Data Preparation

In the context of supervised learning, algorithms acquire knowledge from input data. Therefore, the initial stage of data compilation is of the utmost importance. The data's quality and quantity of features have a significant impact on the efficacy of the models.

Table 2: The Dataset of Credit Card Transactions Used For Research



Feature/s	Type	Description
is_fraud	binary	Determining whether the transaction constitutes fraud or not.
gender	binary	Gender of the customer.
amt	continuous	Amount of the transaction.
city-pop	continuous	Population size of the city the customer resides.
unix-time	continuous	Time of transaction in unix time.
merch-lat / merch-long	continuous	Latitude and Longitude of the Merchant.
lat / long	continuous	Latitude and Longitude of the customer.
street / city / state	nominal	Street, City, and State where customer resides.
zip	nominal	ZIP code on credit card.
first / last	nominal	First and Last name of the customer.
merchant	nominal	Merchant of the customer.
cc-num	nominal	Credit card number of the customer.
trans-num	nominal	Unique transaction number.
category	nominal	Shopping category.
job	nominal	Job of the customer.
trans-day-trans-time	interval-scale	Date and Time of the transaction.
dob	interval-scale	Date of birth of the customer.

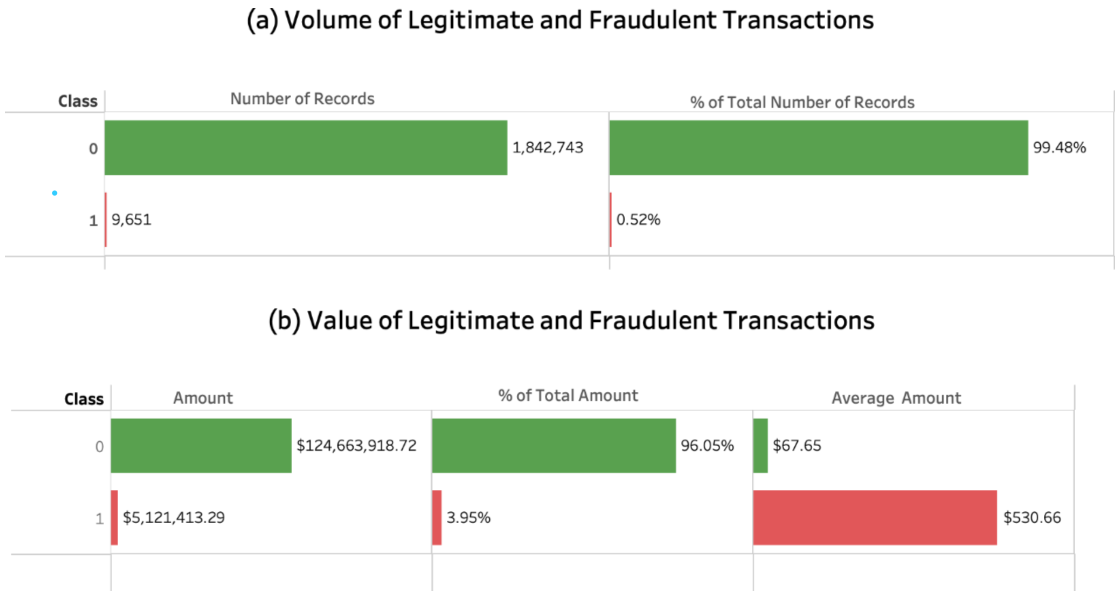


Figure 4 shows the volume and value of both legitimate ('0') and fraudulent ('1') credit card transactions.

Sourced: (Ullastres & Latifi, n.d.-a)

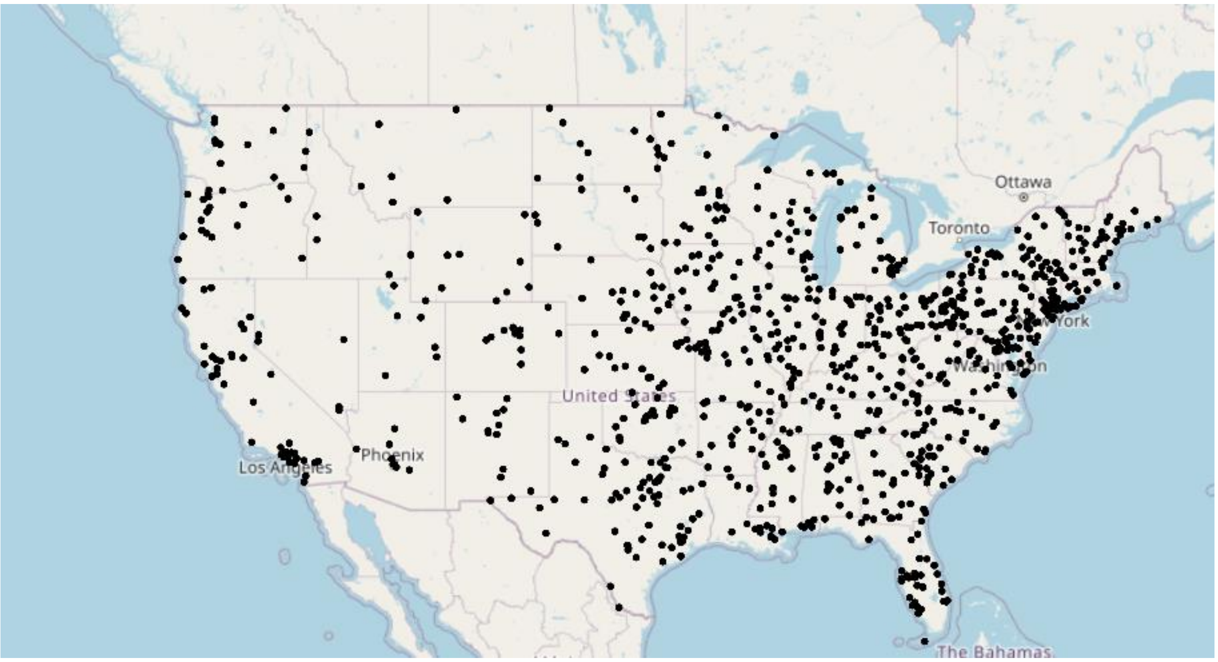


Figure 5: Places with Large Number Fraud Cases

3.3.1. Data Cleaning

To prevent potential overfitting, irrelevant attributes, including unique identifiers, features with nearly unique values, and redundancies, were removed from the dataset. Early in the analysis, attributes such as 'unix-time', 'merch-lat', and 'merch-long' were omitted. After their use in feature engineering, attributes such as 'trans-num' (transaction ID) and 'trans-date-trans-time' were discarded. Noting that the presence of outliers was not addressed is essential. This decision was based on the knowledge that the algorithms utilized in this study are resistant to the impact of outliers.

3.3.2. Data Transformation

The purpose of data transformation is to make data suitable for modelling. During this research analysis, data transformation was performed at two distinct stages. Initially, the binary target variable 'is_fraud' was converted to the data type 'category.' Simultaneously, the attributes 'zip' and 'cc-num' were converted from integers to strings, as these attributes represent nominal variables that are encoded numerically. Additionally, time-sensitive attributes were converted to the datetime format. Another alteration was applied to the 'gender' attribute, which was converted to binary-encoded numbers. Notably, the remaining categorical characteristics, the majority of which have a high cardinality and numerous unique values, have not yet been encoded. The decision to encode them will follow a comprehensive evaluation of their predictive significance in their original format. Subsequently, only the selected categorical features designated for modeling will undergo encoding after a feature selection process.

3.3.3. Feature Engineering

Since the raw formats of the attributes 'trans-date-trans-time' and 'dob' provided limited informational value, feature engineering was performed to extract more insightful attributes. Therefore, the attributes in Table 3 were derived from the 'trans-date-trans-time' variable, while the 'dob' variable contributed to the generation of the 'age' characteristic. As a consequence of this feature engineering procedure, the 'trans-date-trans-time' and 'dob' variables became superfluous and were subsequently removed from the dataset. Likewise, the 'trans-num' attribute was removed because its utility was no longer relevant.

In contrast, the 'region' characteristic was derived from the 'state' variable to provide a location-based descriptor with a reduced cardinality.

3.3.4. Feature Selection

The purpose of feature selection was to improve performance while mitigating overfitting and reducing the training time of the model. Due to the combination of continuous and categorical attributes in the dataset, distinct statistical analyses were conducted to identify the most informative predictors. The correlation matrix was investigated to shed light on the relationship between the continuous predictors and the objective variable. This analysis revealed that the 'amount' predictor, closely followed by the 'hourEncoded' predictor, demonstrated the strongest relationship with the transaction class.

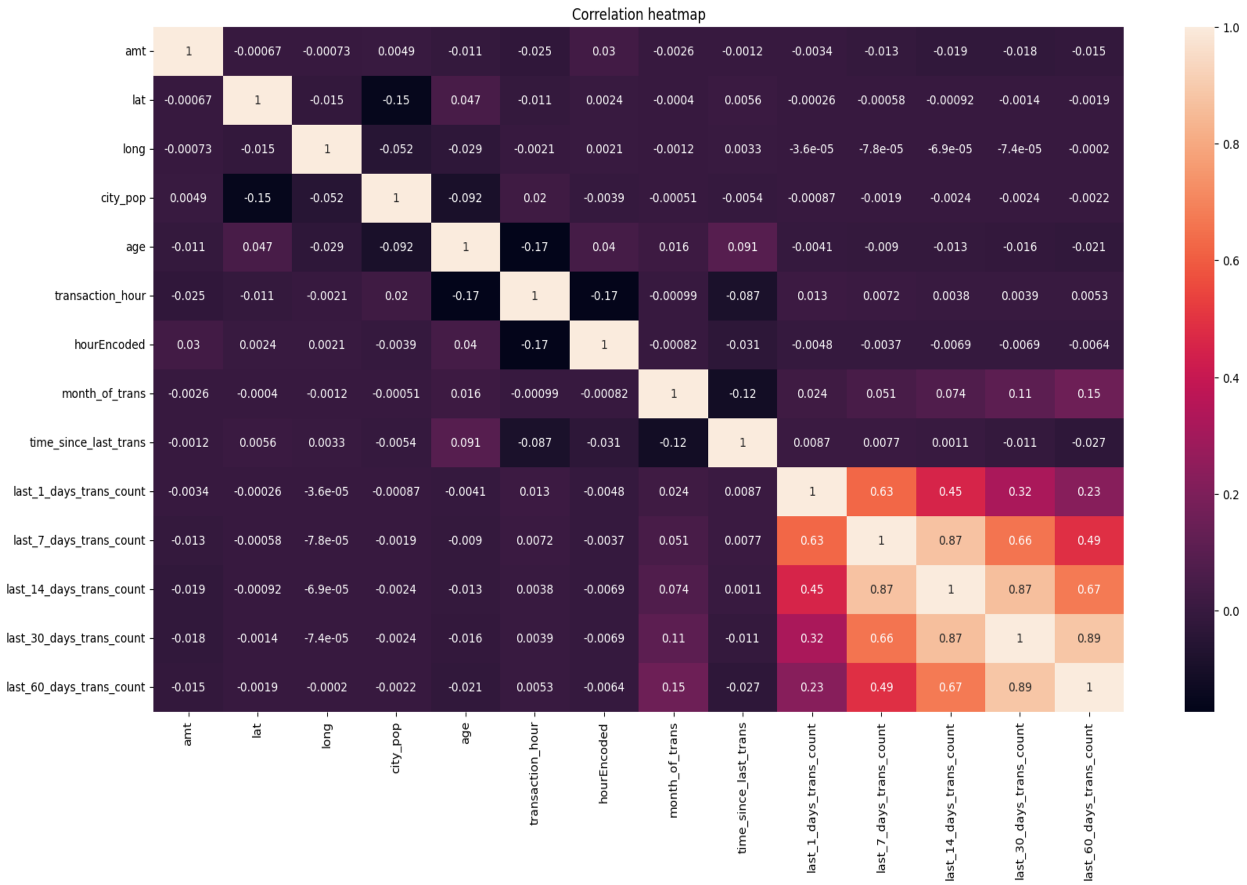
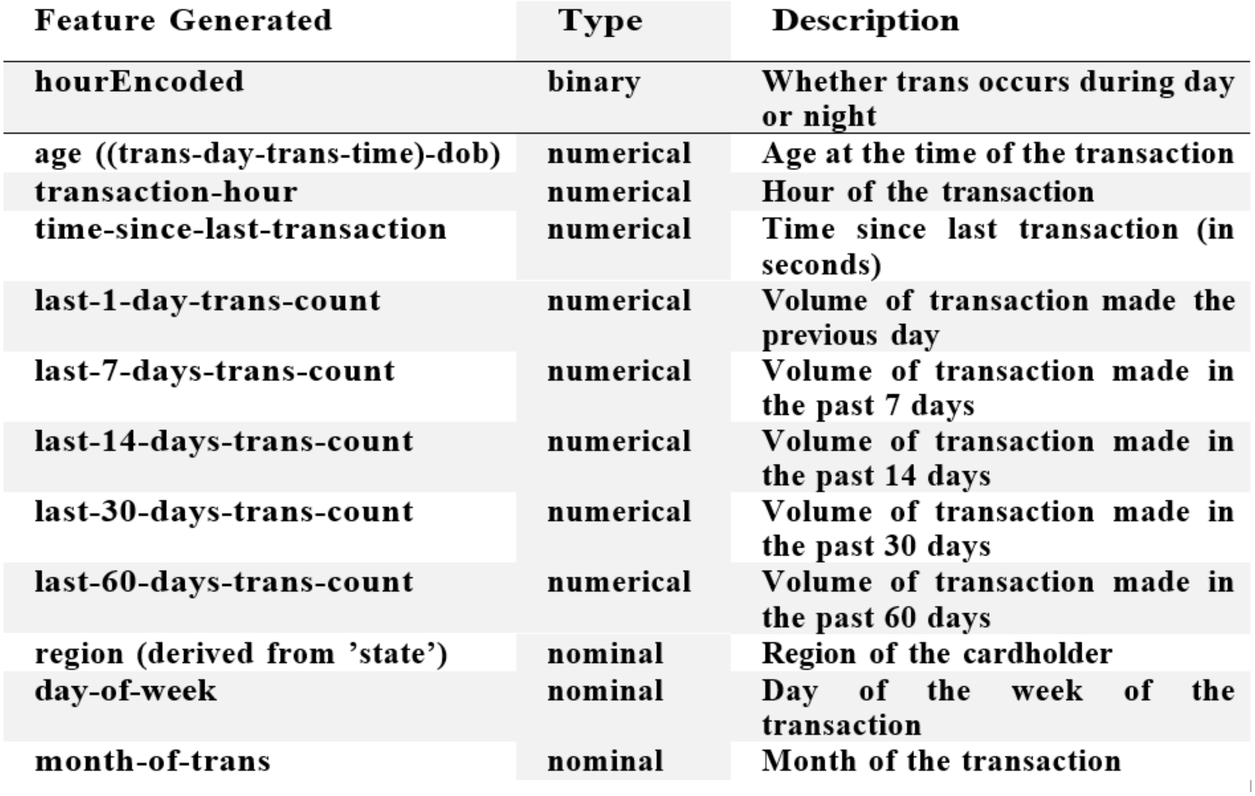


Figure 6: Correlation Heatmap

The univariate method 'SelectKBest' from the 'sklearn.feature_selection' module was used to accomplish feature selection. This technique computes the importance ratings of features using ANOVA F-values (Salekshahrezaee et al., 2023). Out of the initial pool of 15, eight of the most accurate continuous predictors were identified using this method. This choice to employ a filter method is consistent with the objective of this study, as it offers efficiency advantages over wrapper methods, especially for larger datasets.

Table 3: Different Attributes Generated



Feature Generated	Type	Description
hourEncoded	binary	Whether trans occurs during day or night
age ((trans-day-trans-time)-dob)	numerical	Age at the time of the transaction
transaction-hour	numerical	Hour of the transaction
time-since-last-transaction	numerical	Time since last transaction (in seconds)
last-1-day-trans-count	numerical	Volume of transaction made the previous day
last-7-days-trans-count	numerical	Volume of transaction made in the past 7 days
last-14-days-trans-count	numerical	Volume of transaction made in the past 14 days
last-30-days-trans-count	numerical	Volume of transaction made in the past 30 days
last-60-days-trans-count	numerical	Volume of transaction made in the past 60 days
region (derived from ’state’)	nominal	Region of the cardholder
day-of-week	nominal	Day of the week of the transaction
month-of-trans	nominal	Month of the transaction

Adopting a distinct strategy, the selection of the most informative categorical predictors was carried out with care. Given the high cardinality of the majority of categorical variables, model development was made more complex. This compelled a strategic evaluation of the predictive significance of these variables in their unprocessed form prior to their incorporation into the model (Vanhoeyveld et al., 2020).

A Chi-Square test for independence was administered to determine the relationship between categorical predictors and the dichotomous objective variable. Cramer's V test measured the strength of this association. The selection of Chi-Square is motivated by its ability to circumvent assumptions about data distribution. Notably, the Chi-Square statistic is non-parametric and does not assume a normal distribution (medica & 2013, n.d.).

The results revealed that categorical predictors with a high cardinality had a feeble but subtle relationship with the objective variable. Chi-Square testing is characterized by its omnibus nature, making it difficult to identify the specific categories responsible for the association between a multi-level predictor and the target variable. In response, a post-Hoc test with Bonferroni correction was conducted (Lee et al., 2012). This investigation centered on analyzing the relationship between each level of categorical predictors and transaction class. All levels of predictors demonstrated an association with the transaction class, which guided the selection procedure.

Features:- **'amt', 'age', 'hourEncoded', 'time-since-last-transaction', 'last-7-days-trans-count', 'last-14-days-trans-count', 'last-30-days-trans-count', 'last-60-days-trans-count', 'category', and 'day-of-week'** (Ullastres & Latifi, n.d.-a) .

3.3.5. Train and Test Set

To facilitate the training and evaluation of classification models, the dataset was divided into a training set consisting of 70% of the data and a test set consisting of the remaining 30%. Due to the extreme imbalance inherent in the dataset, a stratified dividing strategy was employed with discretion. This method was selected to preserve the proportional representation of classes as observed in the initial dataset. The subsequent construction of the models took place solely within the training set, followed by a thorough evaluation on the specified test set.

3.3.6. Categorical Encoding

In pursuance of effective modelling, a nuanced approach was adopted to incorporate the identified categorical predictors. This innovative strategy involved the use of the CatBoost encoder, a supervised target-based encoding method. This method introduces an inherent ordering principle that reflects the hierarchical structure of categories with respect to the objective variable. The CatBoost encoder's precision-driven statistic computation, which relies solely on the historical target data of each observation, is a defining characteristic. CatBoost encoding is the preferable option for this research endeavor due to its adept management of target leakage concerns, which is a key advantage of this encoding method (de la Bourdonnaye & Daniel, 2021). To ensure the highest level of process integrity and prevent any accidental data leakage, meticulous segregation was maintained. This encoding method was implemented with care to both the training and test datasets.

Furthermore, it is essential to acknowledge that a different method was used to encode categorical predictors. In particular, the widely accepted one-hot encoder method was implemented. This technique transforms categorical variables into binary columns for individual categories, making them suitable as input for machine learning algorithms (Yu et al., 2020). This decision is based on the demonstrated utility of one-hot encoding. Nevertheless, it is essential to approach this decision with caution, taking into consideration the unique characteristics and compatibility with the selected algorithms. This emphasizes the importance of aligning encoding techniques with the distinct characteristics of the dataset and the algorithms in use, a crucial aspect of the data preprocessing pipeline.

3.3.7. Class Imbalance

The dataset under scrutiny within this research demonstrates a pronounced class imbalance, wherein merely 0.52% of transactions manifest as fraudulent instances. This prevailing imbalance introduces a predicament wherein classifiers tend to favor the majority class, thereby potentially yielding an elevated classification accuracy that is, in essence, reflective of the underlying class distribution. Addressing this quandary mandates a nuanced approach to rebalancing the class distribution within the training set. To this end, a hybrid strategy combining random undersampling (RUS) with Borderline-SMOTE oversampling was judiciously applied to the training dataset. It is prudent to note that the test dataset remained unaltered during resampling procedures, as its intended function is solely for model evaluation. Modifying the class distribution of the test set would distort its fidelity to real-world data dynamics.

(Shamsudin et al., n.d.) and (Singh et al., 2022) and (Sisodia et al., n.d.) advocate for the effectiveness of hybrid sampling strategies in conjunction with ensemble classifiers, especially in highly imbalanced fraud datasets. The hybrid strategy combines elements of SMOTE, in which instances of the minority class are replicated, and RUS, which removes instances from the majority class. This composite approach strategically enlarges the representation of fraudulent transactions within the training data, thereby improving the model's ability to recognize complex fraudulent patterns .

Borderline-SMOTE plays a crucial role within the hybrid strategy due to its emphasis on generating synthetic samples close to the decision boundary. This unique strategy promotes the understanding of class boundaries, which is especially important for instances near the class boundary that are more susceptible to misclassification. RUS simultaneously reduces instances from the majority class, ensuring that training data is proportionately distributed (Hajek & Henriques, 2017).

Importantly, the large number of 1,296,675 observations in the training set necessitates prudent management of computational resources. To alleviate the computational burden, a two-pronged approach involving RUS and Borderline-SMOTE was meticulously implemented. These sampling parameters were meticulously fine-tuned with the goal of achieving a balanced majority-to-minority class ratio while preserving a sufficient minority class size. The resulting training set consists of 256,728 transactions, of which 135,120 are legitimate and 121,708 are fraudulent. The implementation of this strategy was facilitated by the imbalanced-learn Python module, which is adept at managing class imbalance scenarios. Figure 5 eloquently illustrates the cumulative effect of the combined RUS and Borderline-SMOTE strategy, demonstrating its capacity to normalize the class distribution within the training dataset.

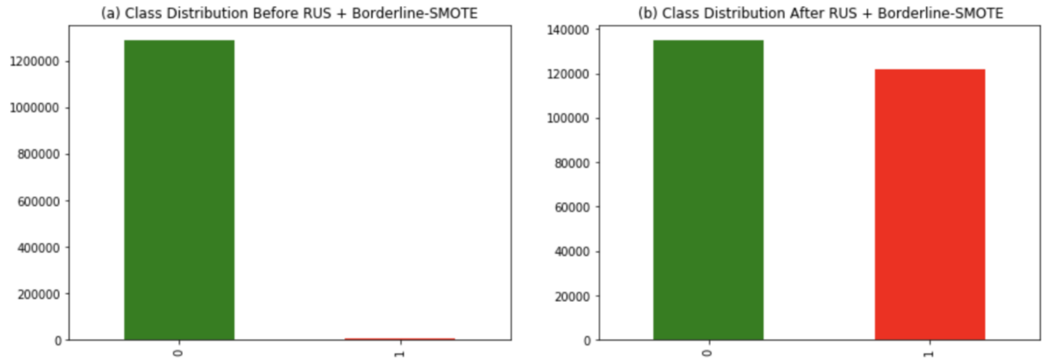


Figure 7: Before and After Applying RUS and Borderline SMOTE

3.4 Data Modelling

The selection of tree-based ensemble learning algorithms was determined by a thorough examination of the relevant literature, which revealed their prominence in the detection of credit card fraud. Ensemble learning methodologies, especially those rooted in decision trees (DTs), have emerged as the leading techniques in this field. This strategy entails combining a large number of machine learning models, often referred to as 'weak' learners, to generate a robust and refined output that improves predictive accuracy. Ensemble learning provides numerous benefits, such as increased model stability and enhanced predictive performance. These advantages are exploited by combining predictions from a variety of foundational models, resulting in an ensemble model identified for its enhanced efficacy and designated as a 'strong' learner.

DTs serve as the foundation upon which these ensemble models are constructed. The effectiveness of DTs lies in their ability to derive elementary decision rules from training data, thereby predicting the class of the target variable. Within the ensemble framework, these DTs serve as foundational learners, from which composite models are constructed (Ileberi et al., 2022). Two prominent ensemble strategies are bagging and boosting, which are notable. Bagging, also known as bootstrap aggregation, reduces the variance of the model's predictions, thereby increasing their robustness. In contrast, boosting serves to reduce bias, thereby enhancing the predictive accuracy of the model. This methodological pair, supported by tree-based ensemble learning, provides the groundwork for the research's subsequent analytical endeavors.

The models used in the research are as shown in the figure 6:

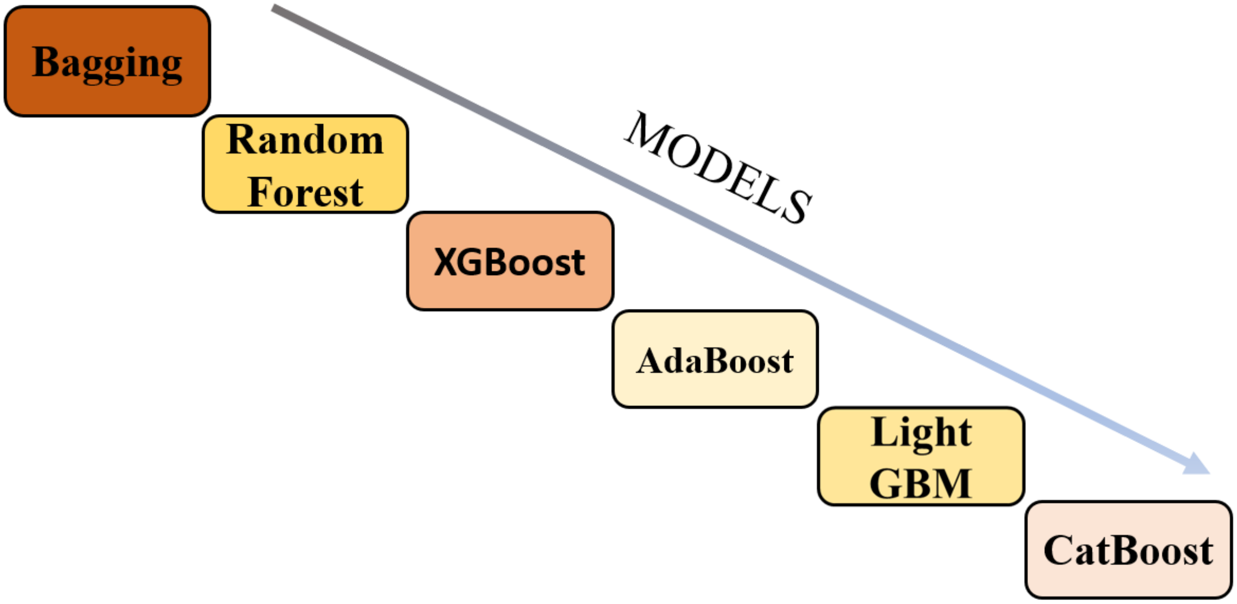


Figure 8: Models That Are Used

3.4.1 Bagging

Bagging uses bootstrap samples, which are random samples with replacement, derived from the training data. On each sample, multiple weak learners, such as Decision Trees (DTs), are simultaneously trained. Through majority voting, the predictions of these individual models are harmonized to produce the final prediction. This combination reduces variance, thereby decreasing the risk of overfitting. Husejinovic et al., (2020) and (Zareapoor et al., n.d.) have demonstrated the effectiveness of bagging in mitigating overfitting, which justifies its use in the present study.

3.4.2 Random Forest

Random Forest (RF) extends Bagging by constructing multiple Decision Trees (DTs) using bootstrap samples from the training data. In contrast to Bagging, Random Forest utilizes a random subset of input features to partition data at each node of the tree. The DT predictions are aggregated, with majority voting favoring the most frequent class. Mishra et al., (2018), Varmedja et al., (2019), Taneja et al., (2019), Muaz et al., (2020), Shamsudin et al., (2020), Sahu et al., (2019), and Dhankhad et al., (2018) selected RF for this study based on its promising performance in previous investigations. Figure 5 visually delineates the structural distinctions between the bagging and boosting methodologies.

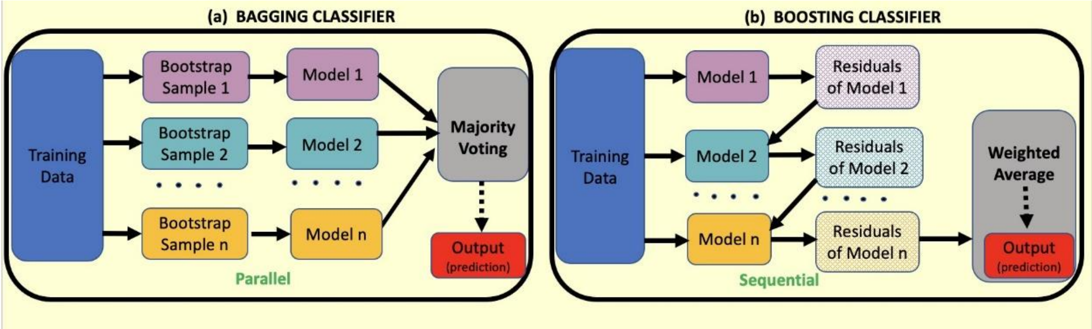


Figure 9: (a) Bagging Classifier. (b) Boosting Classifier.

Sourced: (Ullastres & Latifi, n.d.-b)

3.4.3 XGBoost

Boosting integrates weak learners into a potent model, enhancing predictive efficacy. Unlike bagging, boosting sequentially trains weak learners on multiple weighted training data versions. Each new model improves upon its predecessor's performance by incorporating its residuals. This sequential procedure reduces bias, specifically in Gradient Boosting, which employs gradient descent to refine sequential model precision. XGBoost is a gradient boosting tree algorithm with parallel processing, depth-first tree pruning, integrated cross-validation, and overfitting regularization. Dhankhad et al., (2018); Divakar et al., (2019); Niu et al., (2019) have all demonstrated promising results in fraud prediction using XGBoost.

3.4.4 LightGBM

LightGBM, a gradient boosting tree algorithm, differs from other tree-based methods by employing a vertical leaf-wise growth strategy (best-first) and selecting leaves that minimise expansion loss. This approach, which differs from level-wise growth (depth-first), improves model accuracy, despite the possibility of overfitting due to the complexity of trees. It is essential to calibrate hyperparameters to prevent overfitting. LightGBM outperforms in terms of quicker training periods and decreased memory usage. As demonstrated by Fang et al., (2019) and (Taha & Malebary, 2020), its selection for this study is predicated on its ability to effectively manage large datasets.

3.4.5 CatBoost

CatBoost, a relatively recent gradient boosting tree algorithm, excels at managing categorical features automatically. CatBoost was chosen for this study despite previous categorical feature encoding due to its ability to prevent overfitting and ensure rapid predictions. CatBoost employs symmetric trees for accelerated training and ordered boosting to prevent target leakage and overfitting, unlike other boosting methods. Its superiority over XGBoost in datasets containing mixed data types, as demonstrated by Hancock and Khoshgoftaar (2020), demonstrates its utility.

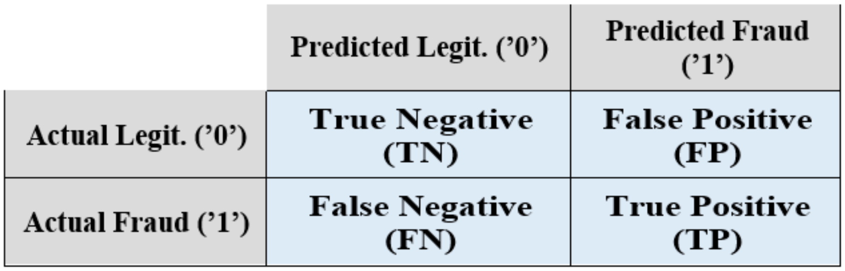
3.4.6 Adaboost

AdaBoost is a well-known algorithm for creating robust classifiers by combining multiple weak classifiers. These weak classifiers are chosen to minimize training errors, resulting in a diverse ensemble of weak classifiers. AdaBoost iteratively modifies sample weights, increasing sample weights for incorrectly classified samples and decreasing sample weights for correctly classified samples. This procedure generates diversity among feeble classifiers, which contributes to AdaBoost's impressive performance. However, its primary focus is error minimization, and it may not always minimize generalization errors (An & Kim, 2010).

3.5 Model Development and Evaluation

Identifying 'fraudulent' and 'legitimate' transactions is the difficulty of credit card fraud detection, which entails a binary classification scenario with unbalanced data. The 'positive' category is the 'fraudulent' category, while the 'negative' category is the 'legitimate' category. The purpose of this study is to develop a model that can effectively identify fraudulent transactions. Recall, Precision, F1-score, Matthews Correlation Coefficient, Geometric Mean, and Area Under the Precision-Recall Curve are used to evaluate the performance of a model based on metrics derived from the confusion matrix, as shown in Table 4.

Table 4: Confusion Matrix for Credit Card Fraud Detection



	Predicted Legit. ('0')	Predicted Fraud ('1')
Actual Legit. ('0')	True Negative (TN)	False Positive (FP)
Actual Fraud ('1')	False Negative (FN)	True Positive (TP)

Recall (also referred to as Sensitivity) (1) quantifies the model's ability to accurately identify fraudulent transactions among all actual fraudulent transactions. This metric is also known as the rate of fraud detection or the true positive rate (Baesens et al., 2015).

Precision (2) measures the model's ability to identify fraudulent transactions among all predicted fraudulent transactions. There is a tradeoff between precision and recall (Baesens et al., 2015).

The F1 score (3) is the harmonic mean of precision and recall. Given the extremely unbalanced nature of fraud datasets, accuracy becomes a deceptive metric for evaluation. The F1-score is a superior performance metric for unbalanced classification challenges, such as credit card fraud detection. In cases of significant data imbalance, however, the F1-score may not provide a comprehensive evaluation because it disregards True Negatives (TNs), thereby producing potentially biassed results (Baesens et al., 2015).

Matthews Correlation Coefficient (MCC) (4) provides an exhaustive evaluation of binary classifications by taking into account the balanced ratio of all four confusion matrix outcomes. This metric reveals the efficacy of the classifier for both positive and negative classes (Chicco & Jurman, 2020).

$$Recall = \frac{TP}{TP + FN} \qquad \ldots\ldots\ldots\ldots\ldots..(1)$$

$$Precision = \frac{TP}{TP+FP} \qquad \ldots\ldots\ldots\ldots\ldots\ldots.(2)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \dots\dots\dots..(3)$$

$$MCC = \frac{[(TP * TN) - (FP * FN)]}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad \dots\dots\dots\dots.(4)$$

$$Geometric\ Mean = \sqrt{Sensitivity * Specificity} \qquad \dots\dots\dots..(5)$$

$$Specificity = \frac{TN}{TN + FP} \qquad \dots\dots\dots\dots\dots(6)$$

The **Geometric Mean (G-Mean)** (5) provides a synthesis of sensitivity (1) and specificity (6) (true negative rate), permitting an evaluation of performance equilibrium across both classification categories. As Tharwat (2020) emphasizes, this metric is especially useful for datasets that exhibit significant imbalances (Tharwat, 2018).

The **Precision-Recall (PR)** Curve depicts the delicate balance between precision and recall across varying thresholds, providing a useful method for evaluating binary classifiers. According to Saito and Rehmsmeier (2015), when evaluating datasets with significant imbalances, the PR Curve is a more informative and precise evaluation instrument than the ROC Curve. The AUC-PR (Area Under the Curve) score quantifies the classifier's ability to effectively discern and differentiate (Saito & Rehmsmeier, 2015).

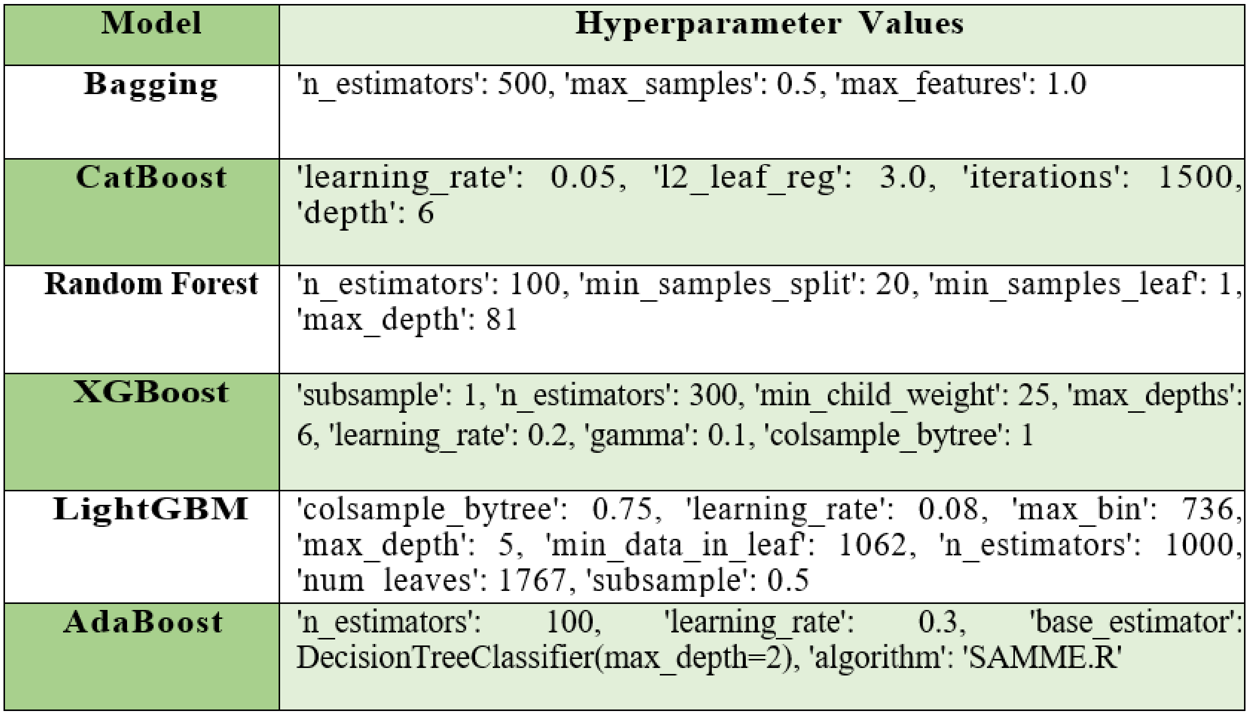
4. Interpretability and Explainability of the Models

The models were developed with the balanced training dataset, as described in subsection 3.3.7. To ensure a more equitable class distribution, the training dataset underwent a hybrid sampling technique involving RUS (Random Undersampling) combined with Borderline-SMOTE (Synthetic Minority Over-sampling Technique). Additionally, the training dataset's size was purposefully downsized, transitioning from the initial 1,296,675 observations to a more manageable scale of 256,728 instances. This modification was made to facilitate the training of multiple models.

The Bagging and Random Forest classifiers were constructed using their respective modules from the Python sklearn.ensemble library. In contrast, the construction of XGBoost, LightGBM, and CatBoost classifiers required the installation of their respective Python libraries using the pip package manager. Subsequently, these libraries were imported to facilitate the model building process. Each classifier was trained using the training set before being applied to the test data in order to make predictions regarding the transaction class.

To improve the predictive capability of these classification models, a hyperparameter optimization procedure was carried out. Randomised Search was utilized to determine the optimal hyperparameter values, thereby optimizing model performance by investigating numerous random permutations within the specified search space. Utilizing the RandomizedSearchCV() function from the sklearn.model_selection library, this optimization procedure was executed independently for each model. This method was favored over grid search due to its computational efficacy, a crucial factor when dealing with large datasets. Given the magnitude of the training set, three-fold cross-validation was utilized to optimize hyperparameters. Table 5 provides a comprehensive summary of the specified hyperparameter values for each model.

Table 5: Models and Hyperparameters



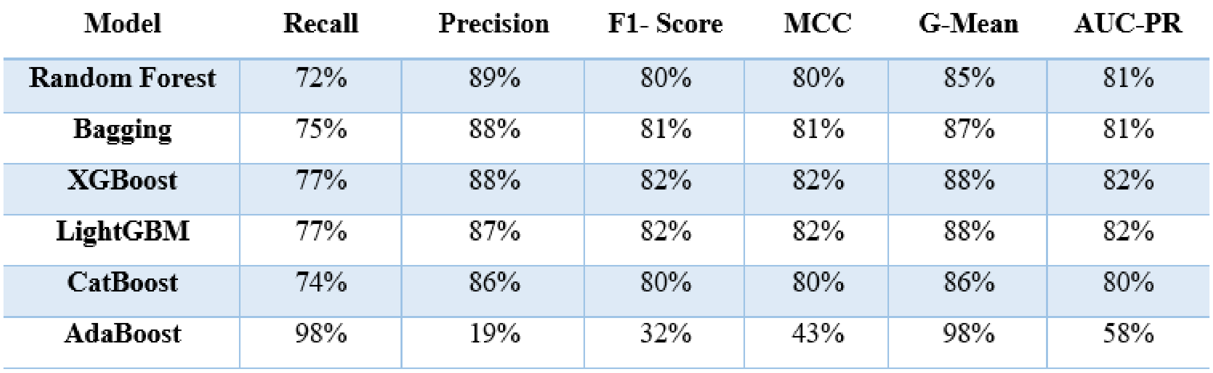
Model	Hyperparameter Values
Bagging	'n_estimators': 500, 'max_samples': 0.5, 'max_features': 1.0
CatBoost	'learning_rate': 0.05, 'l2_leaf_reg': 3.0, 'iterations': 1500, 'depth': 6
Random Forest	'n_estimators': 100, 'min_samples_split': 20, 'min_samples_leaf': 1, 'max_depth': 81
XGBoost	'subsample': 1, 'n_estimators': 300, 'min_child_weight': 25, 'max_depths': 6, 'learning_rate': 0.2, 'gamma': 0.1, 'colsample_bytree': 1
LightGBM	'colsample_bytree': 0.75, 'learning_rate': 0.08, 'max_bin': 736, 'max_depth': 5, 'min_data_in_leaf': 1062, 'n_estimators': 1000, 'num_leaves': 1767, 'subsample': 0.5
AdaBoost	'n_estimators': 100, 'learning_rate': 0.3, 'base_estimator': DecisionTreeClassifier(max_depth=2), 'algorithm': 'SAMME.R'

The optimization process focused on crucial hyperparameters for each model. Regarding the Random Forest model, it is worth mentioning that the default hyperparameter configuration exhibited superior performance compared to the model employing optimized values.

5. Findings

The objective of this project centered on developing a robust model for the precise identification of fraudulent transactions. The evaluation phase consisted of assessing the performance of various classifiers on the test set, which consisted of 555,718 observations, while recognizing the inherent imbalance of the test set, in which fraudulent transactions comprised only 0.52% of the total. Importantly, the class distribution within the test data was not altered, as maintaining a distorted class distribution is consistent with the characteristics of authentic credit card fraud data. The evaluation was conducted using the metrics described in Section 3.5 and the sklearn.metrics module.

Table 6: Accuracy Measures



Model	Recall	Precision	F1- Score	MCC	G-Mean	AUC-PR
Random Forest	72%	89%	80%	80%	85%	81%
Bagging	75%	88%	81%	81%	87%	81%
XGBoost	77%	88%	82%	82%	88%	82%
LightGBM	77%	87%	82%	82%	88%	82%
CatBoost	74%	86%	80%	80%	86%	80%
AdaBoost	98%	19%	32%	43%	98%	58%

5.1 Random Forest

The Random Forest (RF) model exhibits commendable performance in the task of detecting fraudulent transactions. It achieves a Recall score of 0.72, signifying its ability to accurately identify 72% of genuine fraudulent activities, a crucial aspect of risk management in fraud detection. Additionally, RF demonstrates a high Precision level of 0.89, implying that when it designates a transaction as fraudulent, it does so with an impressive accuracy of 89%. This precision substantially reduces the occurrence of false alarms, a pivotal factor in maintaining a low false-positive rate. RF strikes an equilibrium between precision and recall, as indicated by its F1-Score of 0.8, showcasing its effectiveness in correctly identifying fraudulent transactions while maintaining a commendable level of precision. Moreover, the Matthews Correlation Coefficient (MCC) attains a value of 0.8, reflecting the model's robust performance in binary classification tasks, adeptly handling both positive (fraudulent) and negative (legitimate) classes. RF's G-Mean value of 0.85 further underscores its balance in distinguishing between fraudulent and legitimate transactions, especially significant in dealing with unbalanced data scenarios. Furthermore, RF achieves an AUC-PR score of 0.81, highlighting its proficiency in discerning between fraudulent and legitimate transactions across various classification thresholds (Rigatti, 2017). This accomplishment underscores RF's ability to intricately calibrate accuracy and recall, affirming its discriminatory capacity in fraud detection.

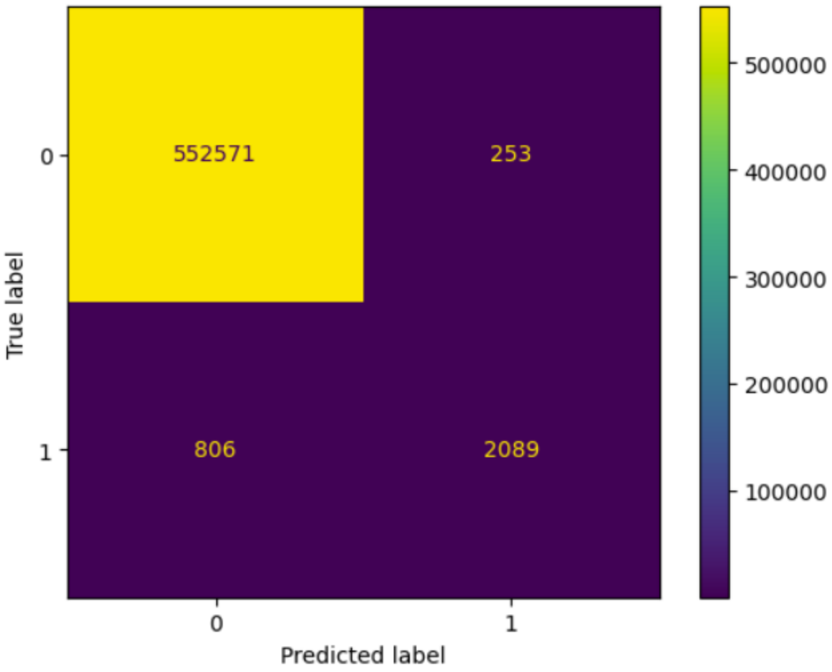


Figure 10: Confusion Matrix of RF

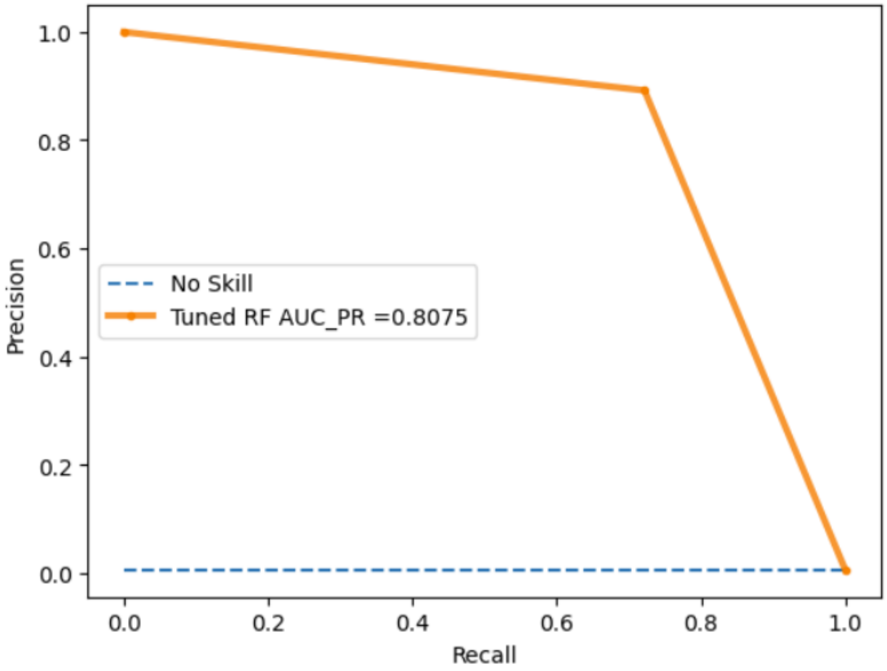


Figure 11: AUC_PR of RF

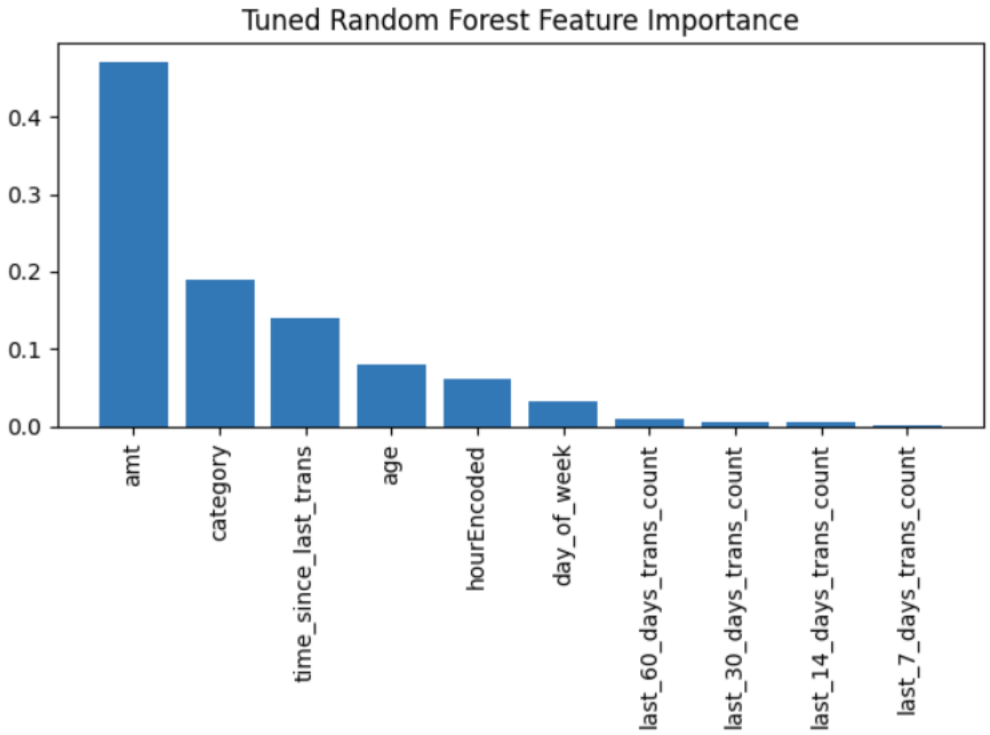


Figure 12: Feature Importance of RF

5.2 Bagging

As a classifier, Bagging exhibits commendable performance in identifying fraudulent transactions. It has a 75% Recall rate, signifying its ability to accurately detect 75% of actual fraudulent activities; this is a significant achievement in fraud detection. In addition, Bagging's Precision score of 88% indicates that when it classifies a transaction as fraudulent, it does so with a high degree of accuracy, resulting in fewer false positives. This level of accuracy is one of the highest among classifiers. Bagging obtains an F1-Score of 0.81, establishing a delicate balance between precision and recall, and identifies fraudulent transactions with commendable precision. The Matthews Correlation Coefficient (MCC) attains a value of 0.81, highlighting Bagging's excellence in binary classification tasks by considering the balanced ratio of true positives, true negatives, false positives, and false negatives. In addition, Bagging receives a G-Mean (Geometric Mean) score of 0.87, indicating a well-balanced approach to differentiating between fraudulent and legitimate transactions, which is particularly useful for managing imbalanced data scenarios. Lastly, Bagging's AUC-PR score of 0.81 demonstrates its efficacy in optimizing precision-recall tradeoffs by emphasizing its ability to differentiate between fraudulent and legitimate transactions across various classification thresholds (Zareapoor et al., 2015).

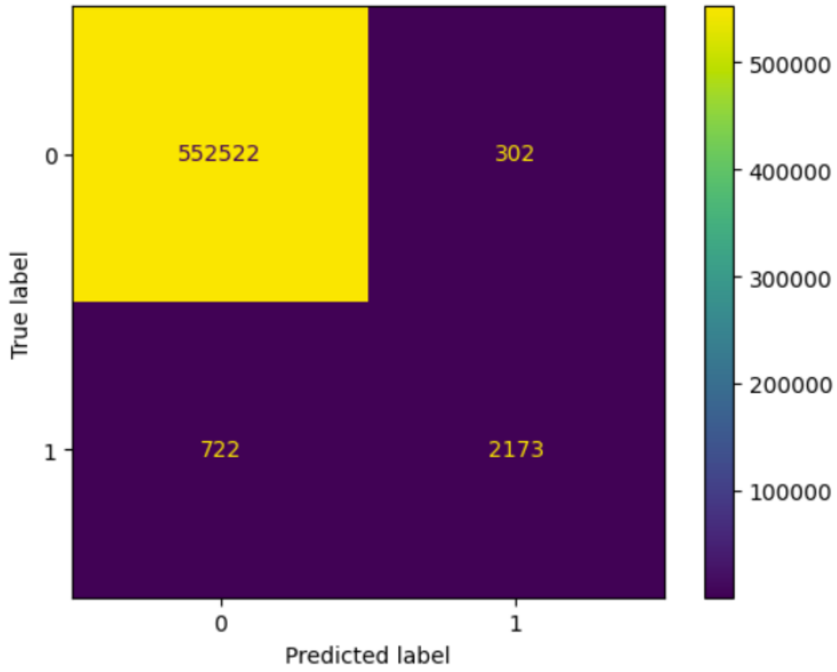


Figure 13: Bagging Confusion Matrix

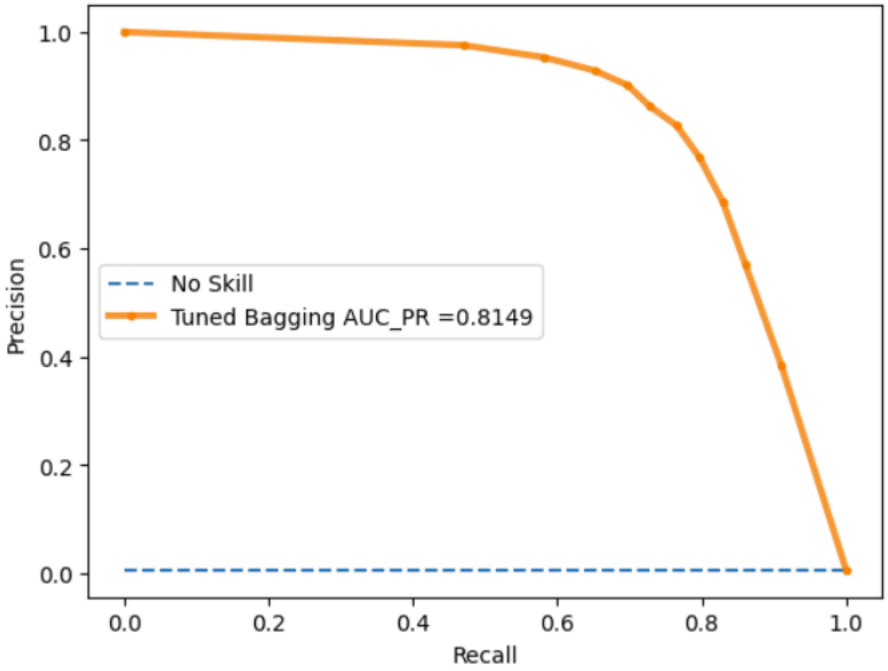


Figure 14: Bagging AUC_PR

5.3 XGBoost

XGBoost emerges as a formidable performer in the detection of credit card fraud. Its significant Recall demonstrates its proficiency in accurately identifying a substantial 77% of actual fraudulent transactions, a crucial attribute for fraud detection. In addition, XGBoost's exceptional Precision of 88% reduces false alarms when flagging transactions as fraudulent, thereby enhancing its dependability (Hancock & Khoshgoftaar, 2020).

The F1-Score of 0.82, which balances Precision and Recall, demonstrates XGBoost's ability to effectively identify fraudulent transactions while maintaining commendable precision, which is essential when dealing with imbalanced datasets. The Matthews Correlation Coefficient (MCC) score of 0.82 demonstrates XGBoost's strong overall performance when positive and negative classes are considered.

Impressively, XGBoost achieves a G-Mean of 88%, demonstrating its accuracy in differentiating between fraudulent and legitimate transactions, thereby addressing the challenges posed by imbalanced data. The AUC-PR score of 0.82 further demonstrates the effectiveness of XGBoost in making informed decisions by optimizing the trade-offs between precision and recall across multiple thresholds. XGBoost is essentially a potent instrument for confronting the complex landscape of credit card fraud detection.

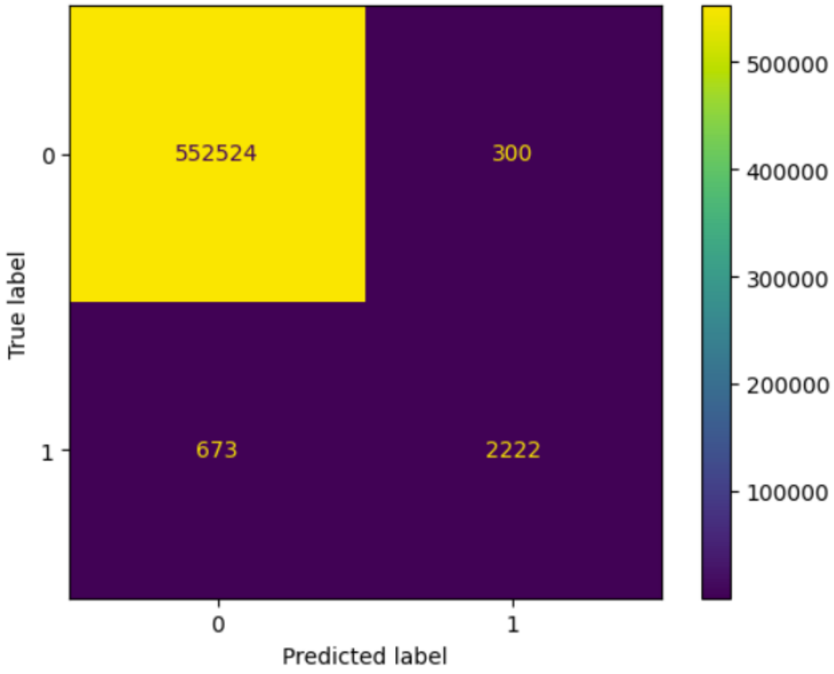


Figure 15: XGBoost Confusion Matrix

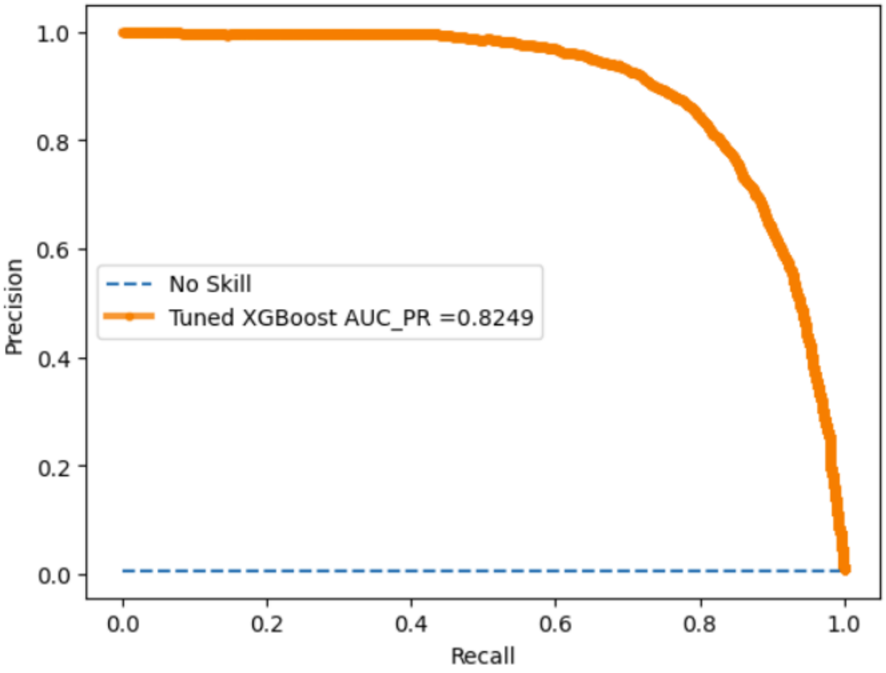


Figure 16: XGBoost AUC_PR

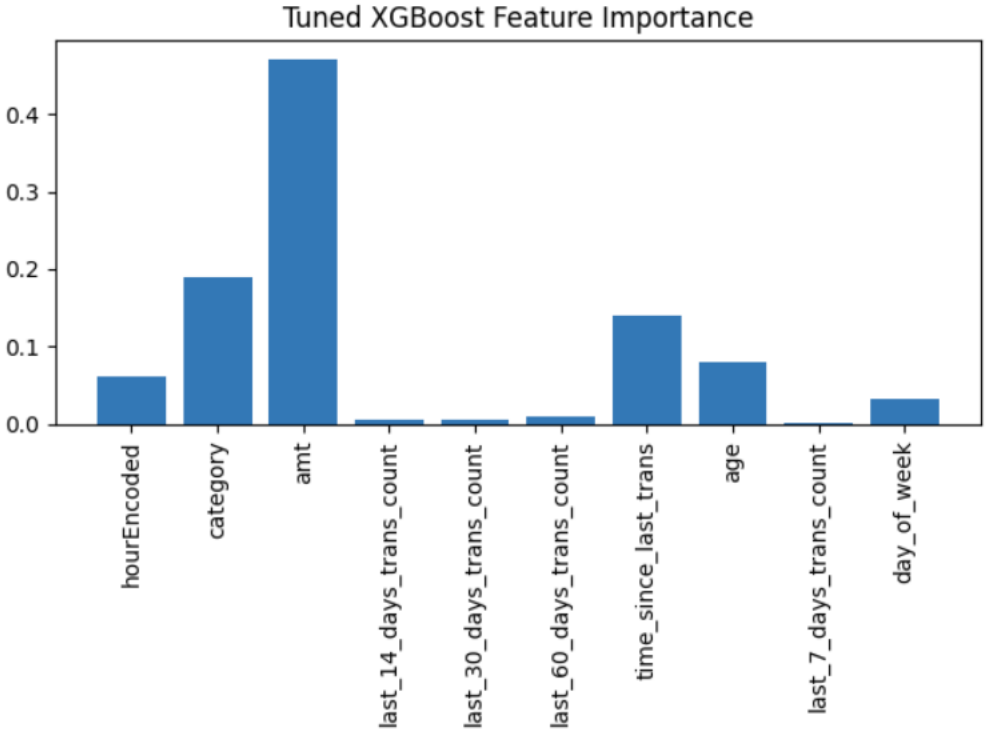


Figure 17: XGBoost Feature Importance

5.4 LightGBM

LightGBM, a robust algorithm in credit card fraud detection, excels across key performance metrics. With a Recall score of 0.77, it demonstrates a commendable ability to identify 77% of actual fraudulent transactions accurately, a vital trait in fraud detection. Moreover, LightGBM exhibits an impressive Precision score of 0.87, indicating an 87% accuracy rate in labeling transactions as fraudulent. This high precision significantly reduces false alarms, crucial for effective fraud detection. Balancing both precision and recall, LightGBM achieves an F1-Score of 0.82, making it adept at identifying fraudulent transactions while maintaining precision, especially important for imbalanced datasets common in fraud detection. The Matthews Correlation Coefficient (MCC) further underscores its excellence, scoring 0.82, reflecting the precision of LightGBM in handling both positive and negative classes. Its high G-Mean of 0.88 signifies balanced accuracy, critical for imbalanced data, and an AUC-PR score of 0.82 highlights its proficiency in optimizing precision-recall trade-offs. Overall, LightGBM's outstanding performance across these metrics makes it a robust and reliable choice for credit card fraud detection, effectively identifying fraud while minimizing false alarms.

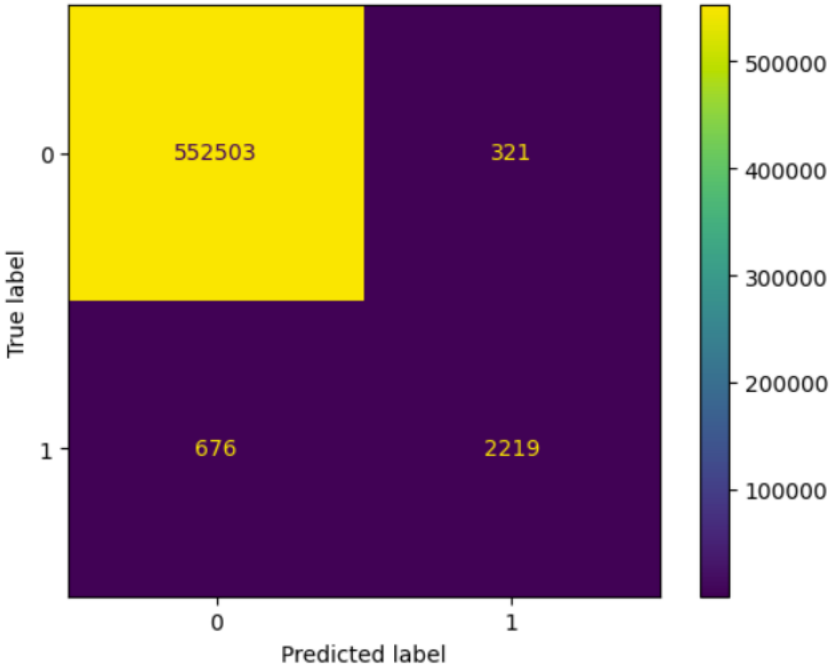


Figure 18: LightGBM Confusion Matrix

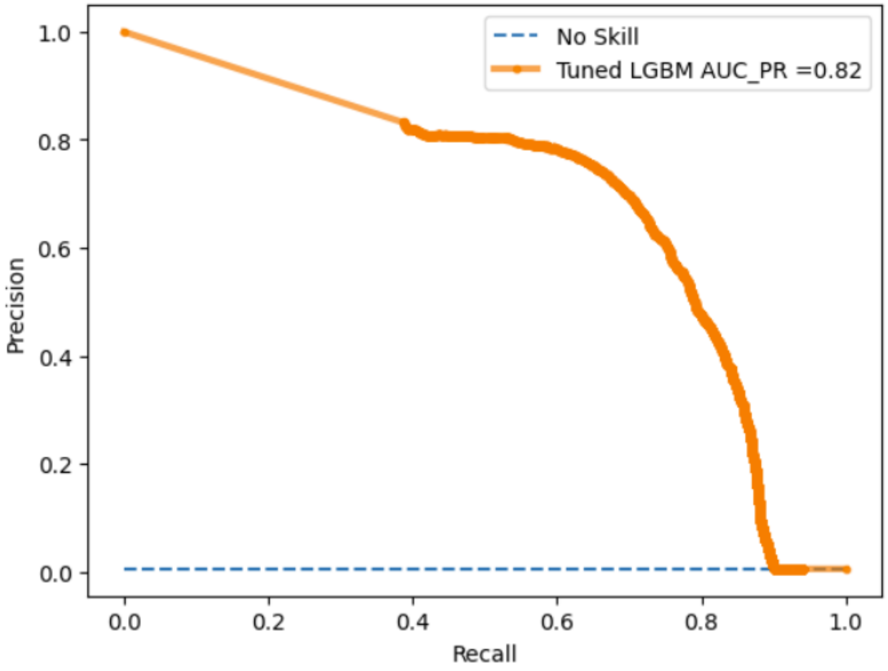


Figure 19: LightGBM AUC_PR

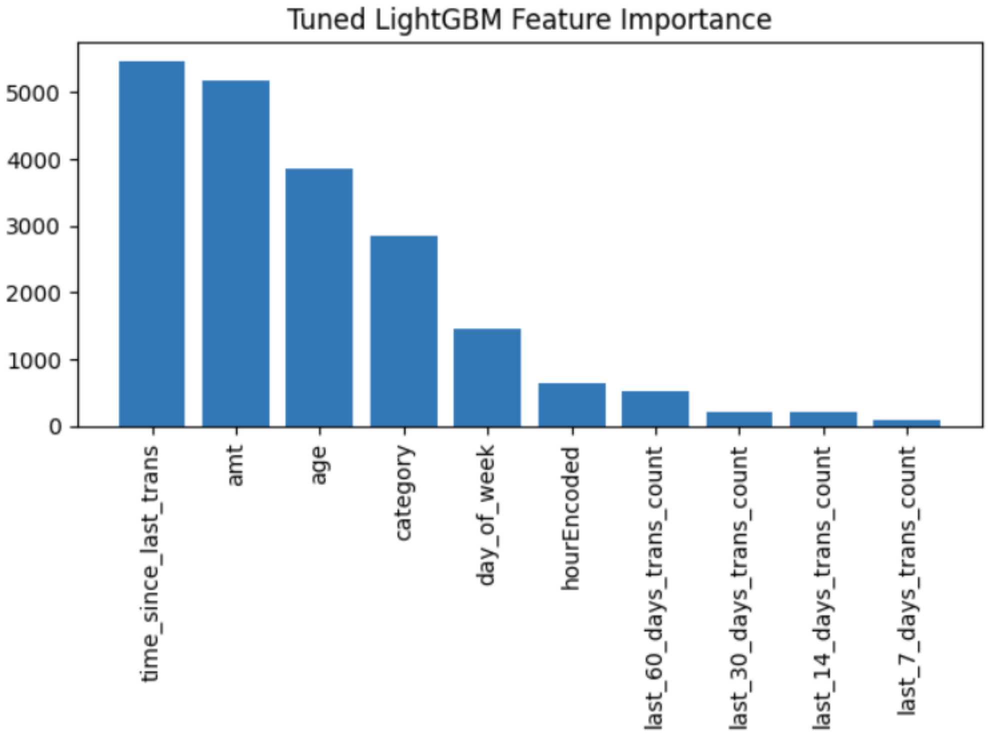


Figure 20: LightGBM Feature Importance

5.5 CatBoost

CatBoost, a formidable competitor in credit card fraud detection, demonstrates commendable capabilities. It accurately identifies 74% of actual fraudulent transactions, as indicated by its Recall score of 0.74, underscoring its efficacy in capturing fraudulent activities. With an impressive Precision score of 0.86, CatBoost excels in labeling fraudulent transactions precisely, achieving an accuracy rate of 86%. This high level of precision significantly reduces false alarms, showcasing CatBoost's reliability in fraud detection. Its F1-Score of 0.8 reflects a balanced fusion of precision and recall, a critical aspect when dealing with imbalanced datasets. The Matthews Correlation Coefficient (MCC) further substantiates its prowess with a score of 0.8, highlighting CatBoost's strength in binary classification tasks by considering the balanced interplay between true positives, true negatives, false positives, and false negatives. CatBoost's G-Mean score of 0.86 signifies balanced accuracy in distinguishing fraudulent from legitimate transactions, while an AUC-PR score of 0.8 underscores its ability to effectively differentiate between these types of transactions across various classification thresholds, optimizing the precision-recall trade-off vital for reliable fraud detection.

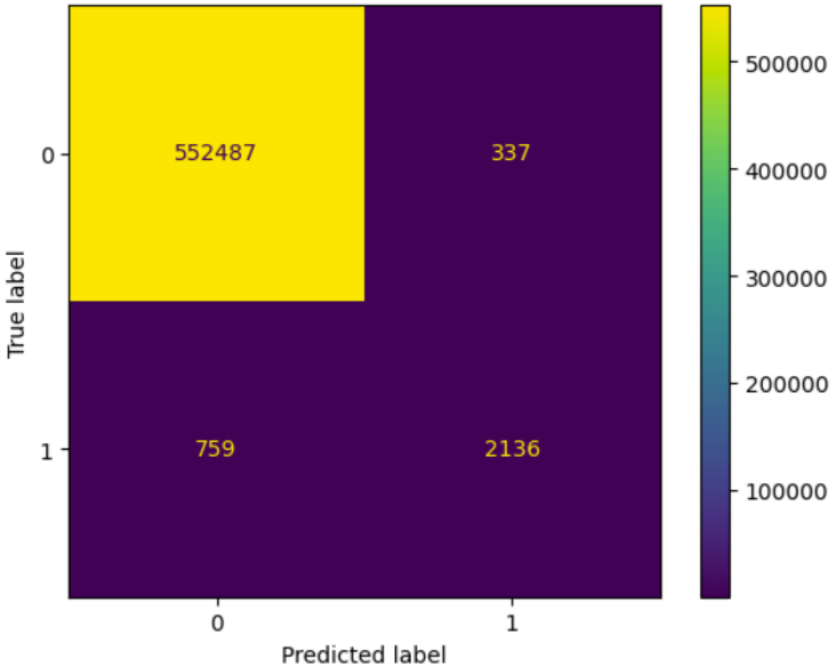


Figure 21: CatBoost Confusion Matrix

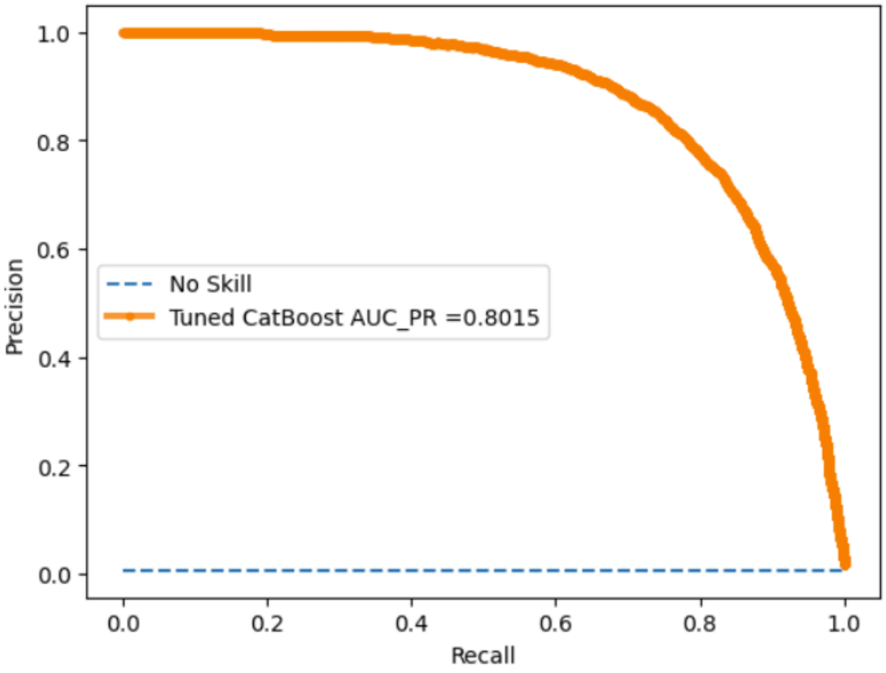


Figure 22: CatBoost AUC_PR

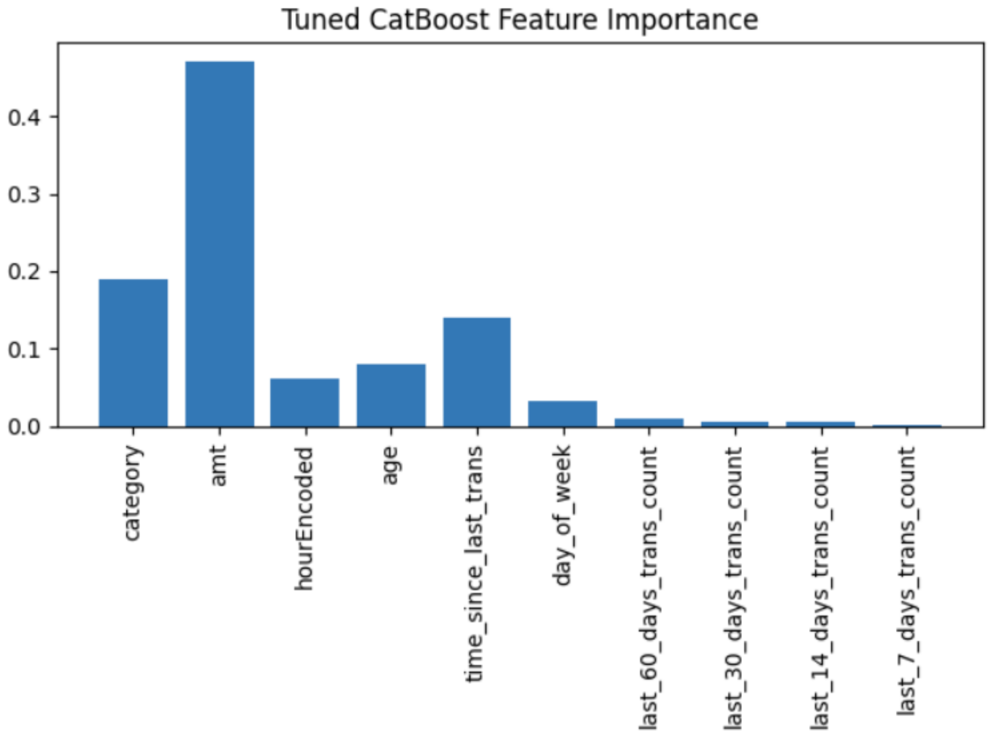


Figure 23: CatBoost Feature Importance

5.6 Adaboost

With a recall of 98%, the AdaBoost model demonstrates a commendable capacity to correctly identify positive cases, indicating its proficiency in capturing nearly all actual positive instances. Nonetheless, it exhibits a comparatively low precision of 19%, indicating that a significant proportion of its positive predictions are false positives, namely 81% of the positive predictions. This reveals a potential development area for reducing the number of incorrect positive identifications. 32% is the F1-score, a proportionate measure of accuracy and recall. The MCC reveals a moderate level of concordance, 43%, between the model's predictions and actual results. Positively, the G-Mean reveals an excellent aggregate performance of 98%. Despite this, the AUC-PR of 58% indicates the possibility for further optimization, especially if a higher precision-recall balance is essential. In conclusion, while the AdaBoost model excels in recall and overall classification, refining precision to reduce the considerable false positive rate could considerably improve its performance, with an emphasis on reducing the 81% rate of false positives to a preferable level (An & Kim, 2010).

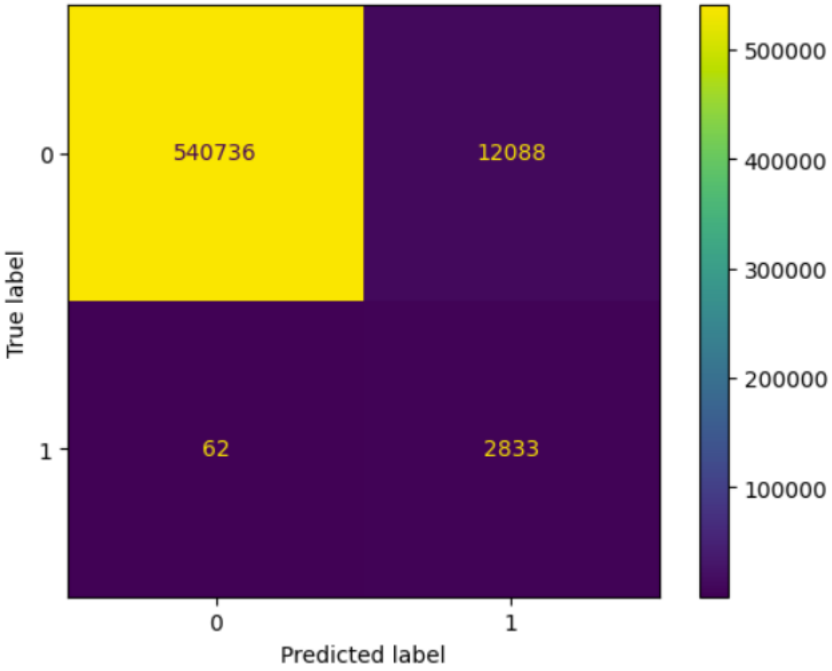


Figure 24: AdaBoost Confusion Matrix

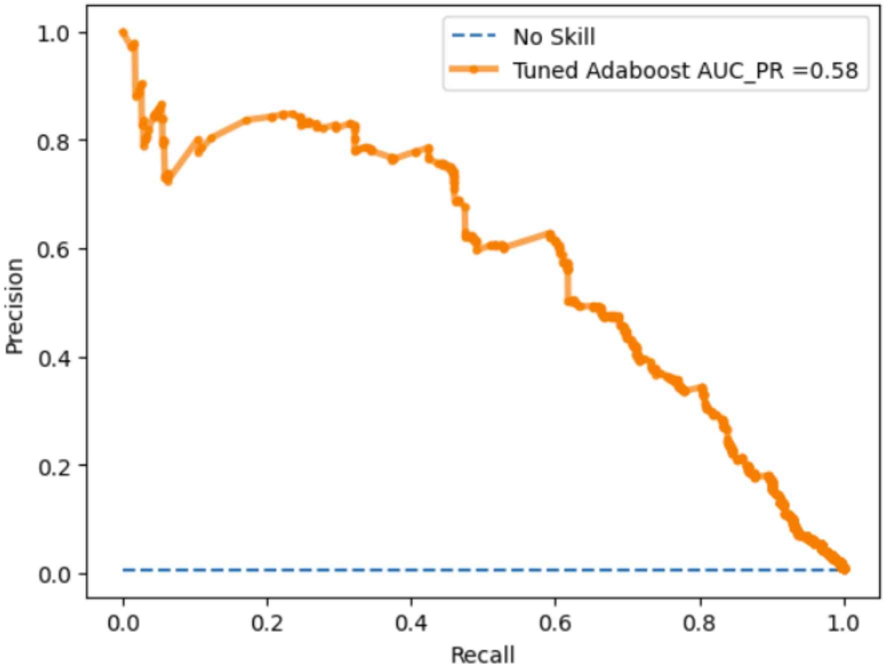
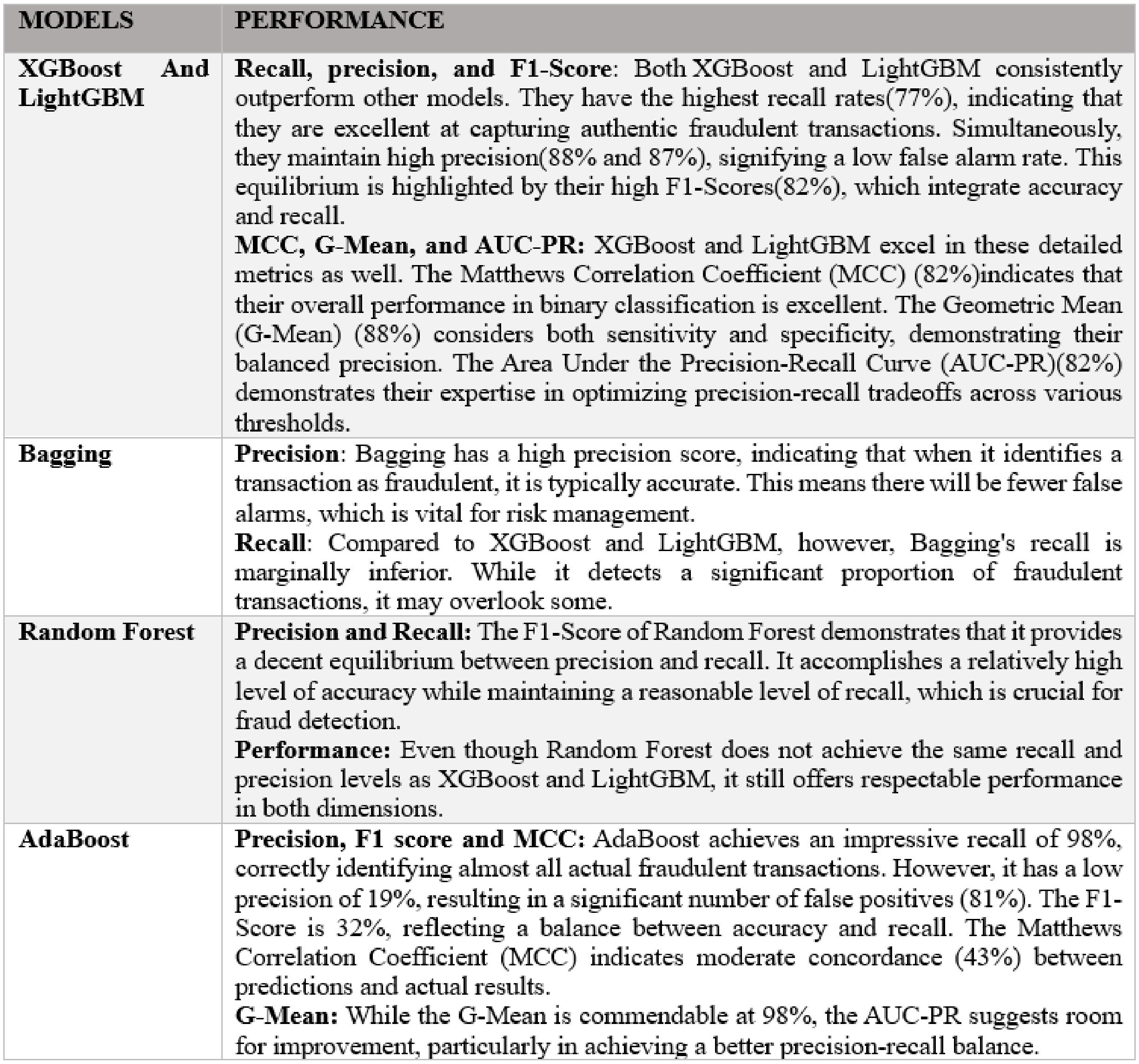


Figure 25: CatBoost AUC_PR

6 Discussion of Findings

The primary objective of this study was to develop a robust model capable of accurately predicting credit card fraud. To accomplish this, sophisticated tree-based ensemble learning algorithms were utilized, with a particular emphasis on bagging and boosting techniques. Notably, the research landscape encircling credit card fraud detection focuses primarily on the use of European datasets. This dataset primarily consists of continuous variables, with extensive anonymization implemented to safeguard sensitive data.

Table 7: Model Comparison



MODELS	PERFORMANCE
XGBoost And LightGBM	**Recall, precision, and F1-Score**: Both XGBoost and LightGBM consistently outperform other models. They have the highest recall rates(77%), indicating that they are excellent at capturing authentic fraudulent transactions. Simultaneously, they maintain high precision(88% and 87%), signifying a low false alarm rate. This equilibrium is highlighted by their high F1-Scores(82%), which integrate accuracy and recall. **MCC, G-Mean, and AUC-PR:** XGBoost and LightGBM excel in these detailed metrics as well. The Matthews Correlation Coefficient (MCC) (82%)indicates that their overall performance in binary classification is excellent. The Geometric Mean (G-Mean) (88%) considers both sensitivity and specificity, demonstrating their balanced precision. The Area Under the Precision-Recall Curve (AUC-PR)(82%) demonstrates their expertise in optimizing precision-recall tradeoffs across various thresholds.
Bagging	**Precision**: Bagging has a high precision score, indicating that when it identifies a transaction as fraudulent, it is typically accurate. This means there will be fewer false alarms, which is vital for risk management. **Recall**: Compared to XGBoost and LightGBM, however, Bagging's recall is marginally inferior. While it detects a significant proportion of fraudulent transactions, it may overlook some.
Random Forest	**Precision and Recall:** The F1-Score of Random Forest demonstrates that it provides a decent equilibrium between precision and recall. It accomplishes a relatively high level of accuracy while maintaining a reasonable level of recall, which is crucial for fraud detection. **Performance:** Even though Random Forest does not achieve the same recall and precision levels as XGBoost and LightGBM, it still offers respectable performance in both dimensions.
AdaBoost	**Precision, F1 score and MCC:** AdaBoost achieves an impressive recall of 98%, correctly identifying almost all actual fraudulent transactions. However, it has a low precision of 19%, resulting in a significant number of false positives (81%). The F1-Score is 32%, reflecting a balance between accuracy and recall. The Matthews Correlation Coefficient (MCC) indicates moderate concordance (43%) between predictions and actual results. **G-Mean:** While the G-Mean is commendable at 98%, the AUC-PR suggests room for improvement, particularly in achieving a better precision-recall balance.

All classifiers attained a high level of specificity, indicating accurate classification of the majority (legitimate) class with a low false positive rate (FPR). Except for Random Forest, most classifiers exhibited high sensitivity, indicating effective fraud detection and low false negatives (low FNR). This led to a high G-Mean score, which measures the equilibrium between sensitivity and specificity (Ghori et al., 2020). Compared to other models, random forest had the lowest G-Mean due to its lesser sensitivity. G-Mean is especially useful for assessing model performance on unbalanced data, which includes both majority and minority class classification.

To enhance the predictive performance of the models, hyperparameter optimization was utilized (Feurer et al., n.d.). Unlike other classifiers, the Random Forest (RF) model performed exceptionally well with its default hyperparameters. For the remaining classifiers, however, fine-tuning their hyperparameters was required and led to significant improvements.

In the context of AdaBoost, the hyperparameters play a crucial role in shaping the behavior and performance of the ensemble classifier. First, the 'n_estimators' parameter is set to 100, indicating that AdaBoost will leverage 100 weak learners, often in the form of decision trees. This ensemble of weak learners collaborates to create a robust and accurate classifier. The 'learning_rate' hyperparameter, set at 0.3, is significant as it governs the influence of each weak learner on the final ensemble. A value of 0.3 suggests a moderate weighting, striking a balance between the contributions of individual learners. The 'base_estimator' parameter is specified as 'DecisionTreeClassifier' with a maximum depth of 2. This designates the type of weak learner employed within AdaBoost, specifically shallow decision trees. Finally, the 'algorithm' is set to 'SAMME.R,' which dictates the algorithm used for weight updating during training. 'SAMME.R' is often considered a superior choice for classification tasks, ensuring efficient ensemble learning. Together, these hyperparameters define AdaBoost's configuration, enabling it to excel in classification tasks, including those related to fraud detection (Feurer et al., n.d.).

For most classifiers, increasing the number of estimators was advantageous up to a certain limit, albeit at the expense of longer training periods. The optimization of the Bagging classifier entailed decreasing the utmost number of features required for training each Decision Tree (DT), resulting in performance improvements.

Boosting classifiers revealed a requirement for a learning rate between 0.2 and 0.3 to obtain commendable results while maintaining reasonable training times. In addition, increasing the minimum number of data points per leaf in the LightGBM model was found to reduce overfitting, thereby improving its efficacy. Setting a maximal tree depth within the ensemble enhanced the performance of both bagging and boosting classifiers, resulting in an overall improvement.

Due to the sizeable training dataset, a three-fold cross-validation method was selected for hyperparameter tuning. In conclusion, boosting models, particularly LightGBM and XGBoost, outperformed bagging models in credit card fraud detection. XGBoost emerged as the favored option, demonstrating marginally superior fraud detection rates and G-Mean values than LightGBM.

7 Conclusion

The study examines the effectiveness of bagging and boosting algorithms for detecting credit card fraud. Using historical transactional data, the plan is to develop a tree-based ensemble classifier capable of accurately distinguishing between legitimate and fraudulent transactions. The employed data set contains both continuous and categorical variables and is characterized by a significant class imbalance. Utilizing feature engineering and selection, it was discovered that high-cardinality demographic attributes have a weak correlation with transaction class. Prior to model development, a hybrid sampling technique incorporating RUS and Borderline SMOTE was applied to the training data to resolve the class imbalance. XGBoost and LightGBM appear to be the most effective models for detecting credit card fraud due to their ability to achieve high recall rates while maintaining high precision. These models strike the delicate balance necessary for accurately identifying fraudulent transactions and minimizing false alarms. However, the selection of the optimal model may also be influenced by the application's specific priorities, such as a focus on accuracy or recall based on business requirements.

The final objective of this research was to pinpoint the most influential predictors of credit card fraud. The analysis revealed that the seven-day transaction volume, the time elapsed since the last transaction, the timing of transactions during the day, transaction quantities, and transaction categories are the primary predictors of credit card fraud.

8 Recommendations

Advanced Anomaly Detection Methods: In conjunction with ensemble classifiers, consider investigating more advanced anomaly detection techniques. Techniques such as autoencoders (Schlegl et al., 2017) and Isolation Forest (Liu et al., n.d.) have demonstrated promise in identifying anomalous credit card transaction patterns. Integrating these methods with ensemble classifiers may enhance the detection of complex fraud patterns.

Real-time Fraud Detection: Investigate the implementation of real-time fraud detection systems capable of analyzing transactions as they occur. Streaming data analytics combined with machine learning can aid in promptly identifying and averting fraudulent activities (Zheng et al., 2010).

Ethical Considerations: As machine learning models become more prevalent in sensitive domains such as fraud detection, it is crucial to investigate the ethical implications. Investigate the potential biases in these models to ensure impartiality and transparency in decision-making (Caton & Haas, 2023).

Feature Engineering: Investigate more advanced feature engineering techniques, such as deep feature synthesis (DFS), utilizing automated machine learning platforms such as Feature tools (Kanarachos et al., 2019). These methods can uncover latent data patterns that may enhance the efficacy of the model (Turner et al., n.d.).

9. Reflective Summary

This machine learning credit card fraud detection research, particularly ensemble approaches like XGBoost and LightGBM, was eye-opening. These methods' accuracy in fraud detection and false alarm reduction are outstanding. Class imbalance must be addressed by hybrid sampling. AI ethics, particularly in sensitive domains like fraud detection, are crucial. Fairness, openness, and privacy are essential. Advanced anomaly identification and real-time analysis offer possibilities for future growth. Ensemble of ensembles brings excitement to model development. This research increased my machine learning understanding and highlighted AI's real-world implications and ethical obligations. It inspires me to study and improve this area.

10 . References

4. Baseline fraud detection system — Reproducible Machine Learning for Credit Card Fraud detection - Practical handbook. (n.d.). Retrieved September 5, 2023, from https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_3_GettingStarted/BaselineModeling.html

An, T. K., & Kim, M. H. (2010). A new Diverse AdaBoost classifier. *Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010*, *1*, 359–363. https://doi.org/10.1109/AICI.2010.82

Arafath, Y., Roy, A. C., Shamim Kaiser, M., & Arefin, M. S. (2022). Developing a Framework for Credit Card Fraud Detection. *Lecture Notes on Data Engineering and Communications Technologies*, *95*, 637–651. https://doi.org/10.1007/978-981-16-6636-0_48

Baesens, B., Vlasselaer, V. Van, & Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. https://books.google.com/books?hl=en&lr=&id=qZwvCgAAQBAJ&oi=fnd&pg=PR15&dq=Baesens,+B.,+Van+Vlasselaer,+V.+and+Verbeke,+W.+(2015).+Fraud+analytics+using+descriptive,+predictive,+and+social+network+techniques:+A+guide+to+data+science+for+fraud+detection,+Hoboken&ots=Ip4juLuN8i&sig=sQ3sPL3KC-U_ebOxg4TJAr6DSdQ

Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*. https://doi.org/10.1145/3616865

Chaudhary, K., Yadav, J., Computer, B. M.-I. J. of, & 2012, undefined. (n.d.). A review of fraud detection techniques: Credit card. *Academia.Edu*. Retrieved September 5, 2023, from https://www.academia.edu/download/74530838/pxc3878991.pdf

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/S12864-019-6413-7

de la Bourdonnaye, F., & Daniel, F. (2021). *Evaluating categorical encoding methods on a real credit card fraud detection database*. https://arxiv.org/abs/2112.12024v1

Dhankhad, S., … E. M.-… international conference, & 2018, undefined. (n.d.). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. *Ieeexplore.Ieee.OrgS Dhankhad, E Mohammed, B Far2018 IEEE International Conference on Information Reuse and, 2018•ieeexplore.Ieee.Org*. Retrieved September 5, 2023, from https://ieeexplore.ieee.org/abstract/document/8424696/?casa_token=Zx28QgqITzQAAAAA:dd2oZe2Q_nB9X5YRvg1UedlyUMpZeWUWXDgYlOtiZcq6BScX2p3SXvhDnkzseb5J2o56H3zfTA

Divakar, K., Eng, K. C.-Int. J. Electron. Commun. Comput., & 2019, undefined. (2019). Performance evaluation of credit card fraud transactions using boosting algorithms. *Ijecce.OrgK Divakar, K ChitharanjanInt. J. Electron. Commun. Comput. Eng. IJECCE, 2019•ijecce.Org*. https://ijecce.org/administrator/components/com_jresearch/files/publications/IJECCE_4356_FINAL.pdf

Fang, Y., Zhang, Y., Computers, C. H.-, Continua, M. &, & 2019, undefined. (2019). Credit Card Fraud Detection Based on Machine Learning. *Cdn.Techscience.CnY Fang, Y Zhang, C HuangComputers, Materials & Continua, 2019•cdn.Techscience.Cn*, *61*(1), 185–195. https://doi.org/10.32604/cmc.2019.06144

Feurer, M., Methods, F. H.-A. machine learning:, & 2019, undefined. (n.d.). Hyperparameter optimization. *Library.Oapen.Org*. Retrieved September 6, 2023, from https://library.oapen.org/bitstream/handle/20.500.12657/23012/1/1007149.pdf#page=15

Ghori, K. M. U., Imran, M., Nawaz, A., Abbasi, R. A., Ullah, A., & Szathmary, L. (2020). Performance analysis of machine learning classifiers for non-technical loss detection. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/S12652-019-01649-9

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152. https://doi.org/10.1016/J.KNOSYS.2017.05.001

Hancock, J., & Khoshgoftaar, T. M. (2020). Performance of CatBoost and XGBoost in Medicare Fraud Detection. *Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020*, 572–579. https://doi.org/10.1109/ICMLA51294.2020.00095

Husejinovic, A. H.-, detection, A. (2020). C. card fraud, & 2020, undefined. (2020). Credit card fraud detection using naive Bayesian and c4. 5 decision tree classifiers. *Papers.Ssrn.ComA HusejinovicHusejinovic, A.(2020). Credit Card Fraud Detection Using Naive, 2020•papers.Ssrn.Com*, *8*(1), 1–5. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521283

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/S40537-022-00573-8

Jain, S. (2019). A comparative analysis of various credit card fraud detection techniques A Multilingual Semantic Web Portal for Unconventional Emergencies View project Leveraging AI in Global Epidemics View project. In *International Journal of Recent Technology and Engineering*. https://www.researchgate.net/publication/332264296

khan, S., . S., Kumar, S., & Kumar, M. H. (2021). Credit Card Fraud Detection Using Machine Learning. *International Journal of Scientific and Research Publications (IJSRP)*, *11*(6), 60–67. https://doi.org/10.29322/ijsrp.11.06.2021.p11410

Lee, J. E. R., Rao, S., Nass, C., Forssell, K., & John, J. M. (2012). When do online shoppers appreciate security enhancement efforts? Effects of financial risk and security level on evaluations of customer authentication. *International Journal of Human-Computer Studies*, *70*(5), 364–376. https://doi.org/10.1016/J.IJHCS.2011.12.002

Liu, F., Ting, K., international, Z. Z.-2008 eighth ieee, & 2008, undefined. (n.d.). Isolation forest. *Ieeexplore.Ieee.OrgFT Liu, KM Ting, ZH Zhou2008 Eighth Ieee International Conference on Data Mining, 2008•ieeexplore.Ieee.Org*. Retrieved September 7, 2023, from https://ieeexplore.ieee.org/abstract/document/4781136/?casa_token=-JnMpi8Nb70AAAAA:qE3u1sUddW3KYsvXLPg_3uhje3Mr7HV09O0xSFjbIXfq9JntTvNgKHb7sxZqfbwcmePBAVwTCA medica, M. M.-B., & 2013, undefined. (n.d.). The chi-square test of independence. *Hrcak.Srce.HrML McHughBiochemia Medica, 2013•hrcak.Srce.Hr*. Retrieved September 5, 2023, from https://hrcak.srce.hr/clanak/152608

Mishra, A., Students', C. G.-2018 I. I., & 2018, undefined. (n.d.). Credit card fraud detection on the skewed data using various classification and ensemble techniques. *Ieeexplore.Ieee.OrgA Mishra, C Ghorpade2018 IEEE International Students' Conference on Electrical, 2018•ieeexplore.Ieee.Org*. Retrieved September 6, 2023, from https://ieeexplore.ieee.org/abstract/document/8546939/?casa_token=tmANidku0WEAAAAA:YIi47Cgjx1O76zsDGQJzWvuSCd4oHPLPyKniYCZW1Z94yZi6XbfT-9oQXcBZuN-bPLJq-WYsQw

Misra, S., Thakur, S., Ghosh, M., & Saha, S. K. (2020). An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction. *Procedia Computer Science*, *167*, 254–262. https://doi.org/10.1016/J.PROCS.2020.03.219

Muaz, A., Jayabalan, M., & Thiruchelvam, V. (2020). A Comparison of Data Sampling Techniques for Credit Card Fraud Detection. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 11, Issue 6). www.ijacsa.thesai.org

Nguyen, T. T., Tahir, H., Abdelrazek, M., & Babar, A. (2020). *Deep Learning Methods for Credit Card Fraud Detection*. http://arxiv.org/abs/2012.03754

nilson report - Google Search. (n.d.). Retrieved September 5, 2023, from https://www.google.com/search?q=nilson+report&rlz=1C1CHBF_enIN917IN917&oq=nilson&gs_lcrp=EgZjaHJvbWUqCQgAECMYJxiKBTIJCAAQIxgnGIoFMg8IARAuGEMYrwEYxwEYigUyEggCEC4YFBiDARiHAhixAxiABDIJCAMQABhDGIoFMgcIBBAuGIAEMg8IBRAuGAoYxwEY0QMYgAQyBwgGEAAYgAQyDQgHEC4YrwEYxwEYgAQyCQgIEAAYChiABNIBCTYyNzVqMGoxNagCALACAA&sourceid=chrome&ie=UTF-8

Niu, X., Wang, L., & Yang, X. (2019). *A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised*. http://arxiv.org/abs/1904.10604

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). *A Comprehensive Survey of Data Mining-based Fraud Detection Research*. https://doi.org/10.1016/j.chb.2012.01.002

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, *47*(1), 31–39. https://doi.org/10.17849/INSM-47-01-31-39.1

Sahu, A., … G. H.-2020 I. 17th I., & 2020, undefined. (n.d.). A dual approach for credit card fraud detection using neural network and data mining techniques. *Ieeexplore.Ieee.OrgA Sahu, GM Harshvardhan, MK Gourisaria2020 IEEE 17th India Council International Conference (INDICON), 2020•ieeexplore.Ieee.Org*. Retrieved September 6, 2023, from https://ieeexplore.ieee.org/abstract/document/9342462/?casa_token=k44PJHIp5tUAAAAA:l1S7iau-p-Y76-4NKB-Xc20P0_AR5GwmwRU6LnIIEbQIG1Xq_BqZfrrCR0ebNl6Pihbqsh-5Cw

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3). https://doi.org/10.1371/JOURNAL.PONE.0118432

Salekshahrezaee, Z., Leevy, J. L., & Khoshgoftaar, T. M. (2023). The effect of feature extraction and data sampling on credit card fraud detection. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/S40537-023-00684-W

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10265 LNCS*, 146–147. https://doi.org/10.1007/978-3-319-59050-9_12

Shamsudin, H., Yusof, U., … A. J.-2020 I. 16th, & 2020, undefined. (n.d.). Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. *Ieeexplore.Ieee.OrgH Shamsudin, UK Yusof, A Jayalakshmi, MNA Khalid2020 IEEE 16th International Conference on Control & Automation (ICCA), 2020•ieeexplore.Ieee.Org*. Retrieved September 5, 2023, from https://ieeexplore.ieee.org/abstract/document/9264517/?casa_token=KD3TnFoG9FkAAAAA:W5FpbJ0NXtCCZ8peyle_FrS0KrZkzwYTe-EgcX-48FksB_3qid__nyaFEZVm3Jt9gDwNTjjAnA

Singh, A., Ranjan, R., … A. T.-E. & T. A., & 2022, undefined. (2022). Credit card fraud detection under extreme imbalanced data: a comparative study of data-level algorithms. *Taylor & FrancisA Singh, RK Ranjan, A TiwariJournal of Experimental & Theoretical Artificial Intelligence, 2022•Taylor & Francis*, *34*(4), 571–598. https://doi.org/10.1080/0952813X.2021.1907795

Sisodia, D., … N. R.-2017 I. I., & 2017, undefined. (n.d.). Performance evaluation of class balancing techniques for credit card fraud detection. *Ieeexplore.Ieee.OrgDS Sisodia, NK Reddy, S Bhandari2017 IEEE International Conference on Power, Control, Signals and, 2017•ieeexplore.Ieee.Org*. Retrieved September 5, 2023, from https://ieeexplore.ieee.org/abstract/document/8392219/?casa_token=WZjZCJBuLUAAAAAA:c5YH0J1ErxzGYlENwU4kgKrCDDMEc86r5mK7Cjiv09xc4av3Dl4UWf6y008WrApnhvkSvT2dNA

Taha, A. A., & Malebary, S. J. (2020). An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. *IEEE Access*, *8*, 25579–25587. https://doi.org/10.1109/ACCESS.2020.2971354

Taneja, S., Suri, B., on, C. K.-2019 I. C., & 2019, undefined. (n.d.). Application of balancing techniques with ensemble approach for credit card fraud detection. *Ieeexplore.Ieee.OrgS Taneja, B Suri, C Kothari2019 International Conference on Computing, Power and, 2019•ieeexplore.Ieee.Org*. Retrieved September 6, 2023, from https://ieeexplore.ieee.org/abstract/document/8940539/?casa_token=Ap07LfeTXBwAAAAA:7Ci3tQLH6vxDcltYJy9GuEZDIPxby0ZCvQe0WbY098YP4U1z3njuAUty1VZp4wJJrh8FMxiigA

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/J.ACI.2018.08.003/FULL/HTML

Turner, C., Fuggetta, A., Lavazza, L., and, A. W.-J. of S., & 1999, undefined. (n.d.). A conceptual basis for feature engineering. *Elsevier*. Retrieved September 7, 2023, from https://www.sciencedirect.com/science/article/pii/S016412129900062X?casa_token=nAy0TuzosZ4AAAAA:kafLotEqWBSPB-UqcmWRHla4ECXQ_Z9Lq_k_Z4ZMd0CUG5X-NWJYoBjPxKJ5Lg5M8MOUc9sTQA

Ullastres, E. F., & Latifi, M. (n.d.-a). *Credit Card Fraud Detection using Ensemble Learning Algorithms MSc Research Project MSc Data Analytics*. https://nilsonreport.com/mention/1313/1link/

Ullastres, E. F., & Latifi, M. (n.d.-b). *Credit Card Fraud Detection using Ensemble Learning Algorithms MSc Research Project MSc Data Analytics*. https://nilsonreport.com/mention/1313/1link/

Vanhoeyveld, J., Martens, D., & Peeters, B. (2020). Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue. *Pattern Analysis and Applications*, *23*(3), 1457–1477. https://doi.org/10.1007/S10044-019-00852-W

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit Card Fraud Detection - Machine Learning methods. *2019 18th International Symposium INFOTEH-JAHORINA, INFOTEH 2019 - Proceedings*. https://doi.org/10.1109/INFOTEH.2019.8717766

Yu, X., Li, X., Dong, Y., & Zheng, R. (2020). A Deep Neural Network Algorithm for Detecting Credit Card Fraud. *Proceedings - 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2020*, 181–183. https://doi.org/10.1109/ICBAIE49996.2020.00045

Zareapoor, M., science, P. S.-P. computer, & 2015, undefined. (n.d.). Application of credit card fraud detection: Based on bagging ensemble classifier. *Cyberleninka.OrgM Zareapoor, P ShamsolmoaliProcedia Computer Science, 2015•cyberleninka.Org*. Retrieved September 5, 2023, from https://cyberleninka.org/article/n/324468.pdf

Zheng, V. W., Zheng, Y., Xie, X., & Yang, Q. (2010). Collaborative location and activity recommendations with GPS history data. *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 1029–1038. https://doi.org/10.1145/1772690.1772795

11. Appendix- 1: Dissertation Checklist Sheet

Name: Moon Karmakar
Date Submitted: 7th September 2023
Signature (Digital): Moon Karmakar
I confirm that my dissertation contains the following prescribed elements:

✓ My dissertation portfolio meets the style requirements set out in the MSc Business Analytics Portfolio Dissertation Handbook including a word count on the front page of each element.

✓ I have reviewed the Turnitin similarity report prior to submission.

✓My dissertation title captures succinctly the focus of my dissertation

✓ My title page is formatted as prescribed in the MSc Business Analytics Portfolio Dissertation Handbook

✓ The abstract provides a clear and succinct overview of my study

✓ Each element contains a Table of Contents, and List of Figures and Tables (where appropriate)

✓ My dissertation contains a statement of acknowledgement (optional)

✓ The Introduction section of the research report, at a minimum, covers each of the following issues:

- Background to/context of the project
- Research question(s), aim(s) and objectives
- Why the project is necessary/important
- A summary of the Methodology
- Outline of the key findings
- Overview of chapter structure of the remainder of dissertation

✓ The Background section of the research report, at a minimum, covers each of the following issues:

- Synthesizes the key technical literature relating to the topic
- Synthesizes the key theoretical literature relating to the topic

✓The methodology section of the research report:

- Contains justification for the tools and method(s) selected
- Details the procedures adopted (e.g. the data source/acquisition, data processing, procedures for maximizing rigor and robustness, methods of data analysis etc.) - Contains ethical considerations and decisions

✓ The findings section of the research report reports the results in detail and provides possible explanations for the various findings

✓ The discussion section of the research report makes appropriate linkages between the findings and the literature reviewed

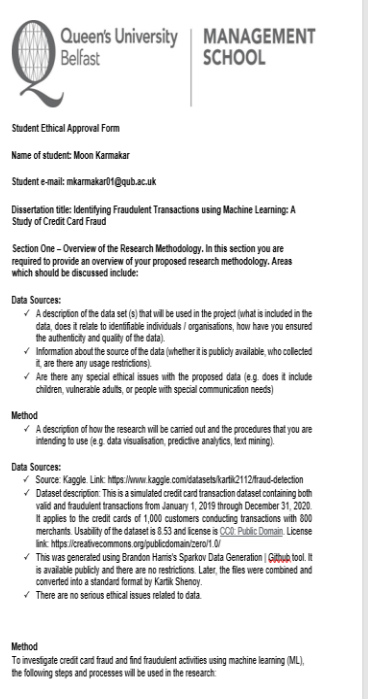
✓ The conclusions section of the research report includes:

- Conclusions about each research question and/or hypothesis
- General conclusions about the research problem
- Implications for theory, for policy and/or management practice
- Limitations of the research
- Suggestions for practice and future research

✓ The technical report, log book, and reflective discussion have each been included.

✓ The reference list is in alphabetical order and follows the Harvard system
✓I have signed and dated the Candidate Declaration

12. Appendix- Ethics Form



Queen's University Belfast | MANAGEMENT SCHOOL

Student Ethical Approval Form

Name of student: Moon Karmakar

Student e-mail: mkarmakar01@qub.ac.uk

Dissertation title: Identifying Fraudulent Transactions using Machine Learning: A Study of Credit Card Fraud

Section One – Overview of the Research Methodology. In this section you are required to provide an overview of your proposed research methodology. Areas which should be discussed include:

Data Sources:

- ✓ A description of the data set (s) that will be used in the project (what is included in the data, does it relate to identifiable individuals / organisations, how have you ensured the authenticity and quality of the data).
- ✓ Information about the source of the data (whether it is publicly available, who collected it, are there any usage restrictions).
- ✓ Are there any special ethical issues with the proposed data (e.g. does it include children, vulnerable adults, or people with special communication needs)

Method

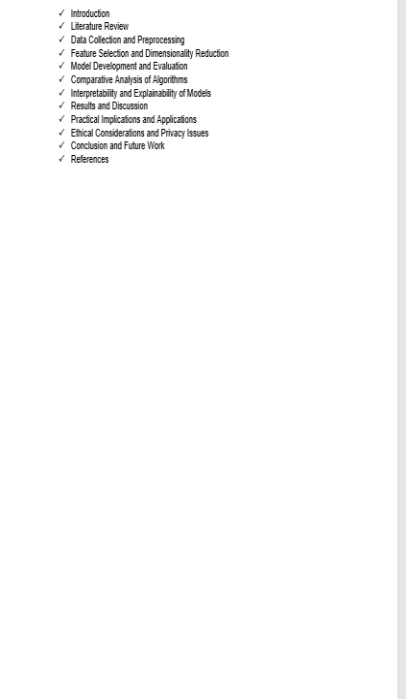
- ✓ A description of how the research will be carried out and the procedures that you are intending to use (e.g. data visualisation, predictive analytics, text mining).

Data Sources:

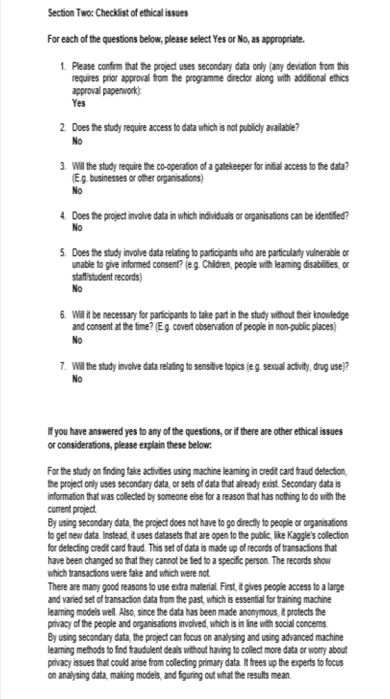
- ✓ Source: Kaggle. Link: https://www.kaggle.com/datasets/kartik2112/fraud-detection
- ✓ Dataset description: This is a simulated credit card transaction dataset containing both valid and fraudulent transactions from January 1, 2019 through December 31, 2020. It applies to the credit cards of 1,000 customers conducting transactions with 800 merchants. Usability of the dataset is 8.53 and license is CC0: Public Domain. License link: https://creativecommons.org/publicdomain/zero/1.0/
- ✓ This was generated using Brandon Harris's Sparkov Data Generation | Github tool. It is available publicly and there are no restrictions. Later, the files were combined and converted into a standard format by Kartik Shenoy.
- ✓ There are no serious ethical issues related to data.

Method

To investigate credit card fraud and find fraudulent activities using machine learning (ML), the following steps and processes will be used in the research:



- ✓ Introduction
- ✓ Literature Review
- ✓ Data Collection and Preprocessing
- ✓ Feature Selection and Dimensionality Reduction
- ✓ Model Development and Evaluation
- ✓ Comparative Analysis of Algorithms
- ✓ Interpretability and Explainability of Models
- ✓ Results and Discussion
- ✓ Practical Implications and Applications
- ✓ Ethical Considerations and Privacy Issues
- ✓ Conclusion and Future Work
- ✓ References



Section Two: Checklist of ethical issues

For each of the questions below, please select Yes or No, as appropriate.

1. Please confirm that the project uses secondary data only (any deviation from this requires prior approval from the programme director along with additional ethics approval paperwork):
   Yes
2. Does the study require access to data which is not publicly available?
   No
3. Will the study require the co-operation of a gatekeeper for initial access to the data? (E.g. businesses or other organisations)
   No
4. Does the project involve data in which individuals or organisations can be identified?
   No
5. Does the study involve data relating to participants who are particularly vulnerable or unable to give informed consent? (e.g. Children, people with learning disabilities, or staff/student records)
   No
6. Will it be necessary for participants to take part in the study without their knowledge and consent at the time? (E.g. covert observation of people in non-public places)
   No
7. Will the study involve data relating to sensitive topics (e.g. sexual activity, drug use)?
   No

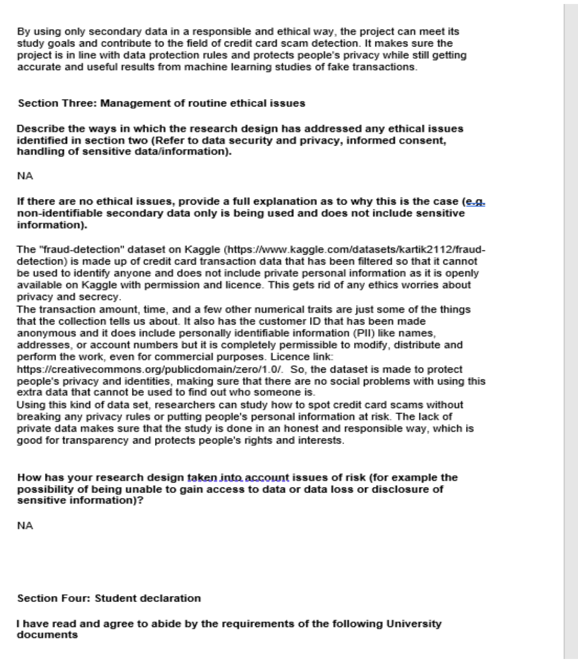
If you have answered yes to any of the questions, or if there are other ethical issues or considerations, please explain these below:

For the study on finding fake activities using machine learning in credit card fraud detection, the project only uses secondary data, or sets of data that already exist. Secondary data is information that was collected by someone else for a reason that has nothing to do with the current project.

By using secondary data, the project does not have to go directly to people or organisations to get new data. Instead, it uses datasets that are open to the public, like Kaggle's collection for detecting credit card fraud. This set of data is made up of records of transactions that have been changed so that they cannot be tied to a specific person. The records show which transactions were fake and which were not.

There are many good reasons to use extra material. First, it gives people access to a large and varied set of transaction data from the past, which is essential for training machine learning models well. Also, since the data has been made anonymous, it protects the privacy of the people and organisations involved, which is in line with social concerns.

By using secondary data, the project can focus on analysing and using advanced machine learning methods to find fraudulent deals without having to collect more data or worry about privacy issues that could arise from collecting primary data. It frees up the experts to focus on analysing data, making models, and figuring out what the results mean.



By using only secondary data in a responsible and ethical way, the project can meet its study goals and contribute to the field of credit card scam detection. It makes sure the project is in line with data protection rules and protects people's privacy while still getting accurate and useful results from machine learning studies of fake transactions.

Section Three: Management of routine ethical issues

Describe the ways in which the research design has addressed any ethical issues identified in section two (Refer to data security and privacy, informed consent, handling of sensitive data/information).

NA

If there are no ethical issues, provide a full explanation as to why this is the case (e.g. non-identifiable secondary data only is being used and does not include sensitive information).

The "fraud-detection" dataset on Kaggle (https://www.kaggle.com/datasets/kartik2112/fraud-detection) is made up of credit card transaction data that has been filtered so that it cannot be used to identify anyone and does not include private personal information as it is openly available on Kaggle with permission and licence. This gets rid of any ethics worries about privacy and secrecy.

The transaction amount, time, and a few other numerical traits are just some of the things that the collection tells us about. It also has the customer ID that has been made anonymous and it does include personally identifiable information (PII) like names, addresses, or account numbers but it is completely permissible to modify, distribute and perform the work, even for commercial purposes. Licence link: https://creativecommons.org/publicdomain/zero/1.0/. So, the dataset is made to protect people's privacy and identities, making sure that there are no social problems with using this extra data that cannot be used to find out who someone is.

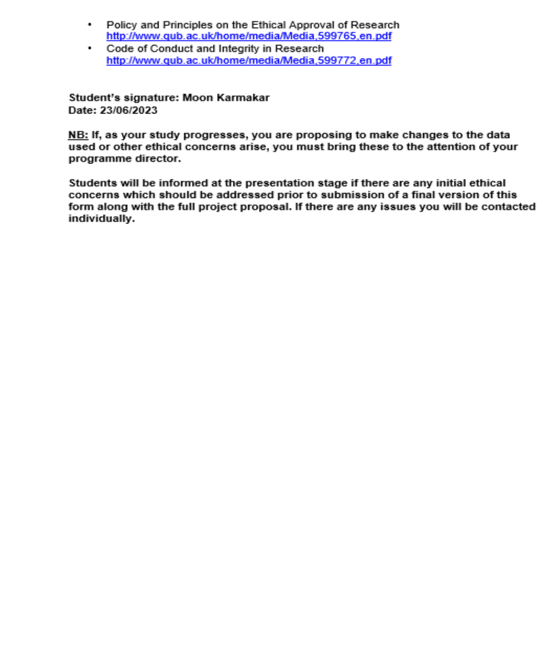
Using this kind of data set, researchers can study how to spot credit card scams without breaking any privacy rules or putting people's personal information at risk. The lack of private data makes sure that the study is done in an honest and responsible way, which is good for transparency and protects people's rights and interests.

How has your research design taken into account issues of risk (for example the possibility of being unable to gain access to data or data loss or disclosure of sensitive information)?

NA

Section Four: Student declaration

I have read and agree to abide by the requirements of the following University documents



- Policy and Principles on the Ethical Approval of Research
  http://www.qub.ac.uk/home/media/Media,599765,en.pdf
- Code of Conduct and Integrity in Research
  http://www.qub.ac.uk/home/media/Media,599772,en.pdf

Student's signature: Moon Karmakar
Date: 23/06/2023

NB: If, as your study progresses, you are proposing to make changes to the data used or other ethical concerns arise, you must bring these to the attention of your programme director.

Students will be informed at the presentation stage if there are any initial ethical concerns which should be addressed prior to submission of a final version of this form along with the full project proposal. If there are any issues you will be contacted individually.