CREDIT CARD FRAUD DETECTION WITH THE USE OF BOOSTING ENSEMBLE MACHINE LEARNING TECHNIQUES

BY

OLUWAPELUMİ ADEYEMİ

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations	Full meanings
CATBOOST	Categorical Boosting
XGBOOST	Extreme Gradient Boosting
GBDT	Gradient Boosting Decision Tree
GBM	Gradient Boosting Model
LGBM	Light Gradient Boosting Model
LightGBM	Light Gradient Boosting Model
BEML	Boosting Ensemble Machine Learning
ADABOOST	Adaptive Boosting
SMOTE	Synthetic Minority Over-sampling Technique
SMOTE-ENN	Synthetic Minority Over-sampling Technique with Edited Nearest Neighbours
CNP	Card Not Present
SVM	Support Vector Machine
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
ADASYN	Adaptive Synthetic Sampling
SHAP	SHapley Additive exPlanations
CCFD	Credit Card Fraud Detection
ML	Machine Learning
CNR	Card Not Received

ABSTRACT

Credit card fraud remains a pervasive challenge in the financial industry, necessitating the development of robust and efficient fraud detection systems. This research project focuses on leveraging boosting ensemble machine learning algorithms to improve the accuracy and reliability of credit card fraud detection.

The study commences with the exploration of various boosting ensemble methods, including Adaboost and Gradient Boosting Decision Trees (GBDT) models namely, GBM, XGBoost, LightGBM, and CatBoost. These techniques are evaluated on multiple datasets, each addressing the class imbalance inherent in fraud detection tasks. Data preprocessing techniques, including resampling methods such as Random undersampling, SMOTE oversampling and SMOTE-ENN resampling, are employed to tackle class imbalance. The performance of these methods is assessed using key metrics like accuracy, precision, recall, and F1-score.

The research culminates in a comprehensive evaluation of the models, comparing their performance across various datasets and resampling techniques. Notably, the results reveal that Gradient Bosting Decision Trees (GBDT) particularly XGBoost and CatBoost models consistently outperform other boosting methods, emphasizing their efficacy in capturing intricate fraud patterns. The findings of this research project contribute to the advancement of credit card fraud detection systems, offering insights into the selection of optimal ensemble methods, resampling strategies, and interpretability techniques. These insights can empower financial institutions to enhance their fraud detection capabilities and bolster security measures in an ever-evolving digital landscape.

1. INTRODUCTION

Our daily lives have become intertwined with digital transactions, the use of electronic payments and online purchases have increased rapidly, leading to an exponential rise in the use of credit cards. This increase in digital transactions has inevitably led to a rise in fraudulent credit card activity as criminals constantly evolve their methods to stay ahead of security measures, causing massive financial losses for both financial institutions and customers.

There is a thin line between fraudulent and lawful behaviour, this makes it difficult to distinctly define fraud. The most vulnerable party is frequently a customer or a merchant, but fraud is very versatile and thrives in situations where it can be very successful.

Nguyen (2014) asserts that all kinds of fraud share these four elements: an intentional attempt to deceive, a materially false statement, knowledge that the assertion is false, and the harm or loss brought on by the deception. The loss resulting from fraud in the United Kingdom, for the year 2022 alone is over £1.2 billion, out of which payment cards had over two million cases leading to a loss of about £556.3 million. This kind of loss could pose a huge impact to consumer spending patterns, financial institution profitability and have a deteriorating effect on the overall economic stability of the country.

FRAUD TYPE	TOTAL LOSSES IN 2022	YEAR ON YEAR CHANGE FROM 2021	TOTAL NUMBER OF CASES IN 2022	YEAR ON YEAR CHANGE FROM 2021
Payment Cards	£556.3m	6%	2,732,894	-3%
Remote Banking	£163.1m	-18%	47,473	-46%
Cheque	£7.5 m	+18%	966	+19%
Total	£726.9m	0%	2,781,333	-5%

Figure 1: Source – UK Finance, 2022

Credit cards are regarded as "an appealing object of fraudulent activities" since attackers may obtain large sums of money in a short period of time with little risk, and nearly all the time the fraudulent activity is found within a few days. TransUnion UK, a global credit-reporting agency defines Fraud as any dishonest or deceptive activity carried out with intent to gain unauthorized access to someone's credit card information or to make fraudulent transactions. They further explained that this involves the use of stolen or counterfeit card details, unauthorized use of

someone else's credit card, or manipulation of the payment process for personal gain. (TransUnion UK, No date)

Credit Card Fraud is described as when a person uses another person's credit card for private purposes while both the owner of the credit card and the company that issued it are unaware that the card has been used. Furthermore, the person using the credit card has no link to the cardholder or the provider and has no plan of contacting the card's owner or repaying for the transactions made. (Bhatla, et al., 2003)

Several studies have tried to class card fraud into different categories. Zhang, et al. (2021) states that credit card fraud is divided into application fraud and behaviour fraud which mainly includes theft/stolen-card fraud, counterfeit-card fraud, and card-not-present fraud. They further explained that the most common type of fraud is theft/stolen card fraud where fraudsters aim to spend as much as they can after stealing a credit card or retrieving a lost card.

Bhatla, et al. (2003) classified card related fraud into application fraud, lost/stolen card, account takeover and counterfeit cards. The study shows that although lost/stolen card is the most common type of fraud, there is an increasing threat of card not present fraud, as the world is drifting into online transactions.

The Covid-19 pandemic in 2020, brought about an overwhelming increase in digital transactions, with the growth of e-commerce, mobile payments, and application subscriptions. This has made it possible for fraudsters to develop sophisticated ways by which they can successfully execute the card not present fraud making it the popular type of credit card fraud presently, other types include counterfeit cards, identity theft, lost/stolen cards and card not received fraud. The UK finance annual report shows that card not present fraud has about 81% contribution to the number of card frauds in the year 2022.

Figure 2: Source- (UK Finance, 2023)

A brief explanation of these types of card fraud is as seen below:

- **Card not present (CNP) fraud:** This type of card fraud is also known as remote purchase fraud. It occurs when the fraudsters use a stolen credit card to make online purchases where the physical card is not present, in most cases, these card details are gotten through hacking or data breaches via phishing emails and scam text messages. Due to its versatility, CNP fraud is a major concern in the e-commerce industry, and various technologies such as secure payment gateways, tokenization, and two-factor authentication, are employed to combat this type of fraud. The UK Finance report states that CNP fraud is about £395 million in 2022 which is a 4% decrease from the previous year, this shows that there has been an improvement in methods by which these kinds of fraud can be detected over the years. On the contrary, Insider Intelligence states that by 2024, 74% of losses will be attributed to card not present fraud which suggest that there is need for development of more accurate and efficient fraud detection models, in a bid to helping organizations and stakeholders protect their financial resources.
- **Counterfeit Cards**: Counterfeit cards are created by copying the information on a legitimate credit card onto a fake card, and then the fake card is used for unauthorized transactions. According to Jha & Westland (2013), offenders of counterfeit card theft steal credit card information to construct an actual counterfeit card. These cards are used to conduct fraud without the real cardholder's awareness. Cardholders may only discover fraud by monitoring credit card usage and reporting any fraudulent activity to the issuing bank or other applicable authorities. According to the UK Finance report, counterfeit card losses amounted to £4.7 million in 2022, which is 97% less than the category's 2008 peak. The implementation of chip technology in the UK, followed by its widespread acceptance, has led to a dramatic drop in the number of counterfeit cards being used by fraudsters.
- **Lost/Stolen Card**: Lost or stolen card fraud occurs when unauthorized individuals use the lost credit card to make purchases before the cardholder can report the loss or theft. In the case that a card is stolen or lost, according to Jha & Westland (2013), fraudsters try to commit as many frauds as they can before the crime is reported and the cards are deactivated. The possibility of minimising fraud increases with the speed at which the card theft is discovered. The UK Finance report states that losses resulting from lost and stolen credit card theft climbed by 30% in 2022 and totalled £100.2 million, the first time this category's losses have topped £100 million on record. To protect cardholders from this kind of fraud, there is an inbuilt security feature that request for their PIN to prove possession and ownership of the card. Various range of fraud prevention and detection tools have been deployed to protect consumers, merchants and financial institutions form this kind of fraud.
- **Card Identity Theft**: This involves stealing someone’s personal information, including credit card details to make unauthorized transactions or open a fraudulent account in

the owner's name. Third-party application fraud and takeover of accounts fraud are two subtypes of this fraud (UK Finance, 2023). The third-party application fraud includes opening a card account in someone else's name using stolen or forged papers. The victim's information is often obtained through data loss, such as through data thefts and using social engineering to breach private information. On the other hand, account takeover fraud refers to a type of card identity theft where a scammer manages to sneak into someone's card account without their permission. A total of £51.7 million was recorded as losses from ID card theft which is a 97% increase from 2021.

- **Card not received (CNR) fraud**: This is also known as "non-receipt fraud". It occurs when a legitimate credit card is issued by the card issuer but is stolen during the delivery process before reaching the intended cardholder, the fraudster gains unauthorized access to the card and uses it for fraudulent purposes. The UK Finance study claims that to conduct this kind of fraud, fraudsters often target buildings with shared letterboxes, such as apartments, dorms for students, and outdoor mailboxes. In 2022, losses from non-receipt fraud were £4 million, an increase of 1% from the year before.

1.1. RESEARCH PROBLEM

One of the major issues in combating credit card fraud is that it impacts multiple stakeholders, including but not limited to financial institutions and merchants. With the growth of online transactions, e-commerce merchants have also been impacted by credit card fraud as they are vulnerable to chargebacks, which involves the reversal of a transaction due to fraudulent activity. Most credit card transaction records include both category and numerical variable, 'transaction amount' is a common example of a numerical feature, while categorical characteristics include things like the merchant's code, name, and date of the transaction. Several of these categorical variables may contain hundreds of thousands of categories, depending on the dataset. This combination of a few discrete numerical qualities and a lot of categorical attributes has led to the employment of several machine learning, statistical, and data mining technologies for detection of fraudulent credit card transactions. (Bhattacharyya, et al., 2011)

Since balanced datasets are incredibly uncommon in real-world problems, Alfaiz & Fati (2022) found that traditional machine learning classification method frequently diminishes the significance of the minority class in the dataset. This minority class should be given the utmost importance since it contains the fraudulent instances, which constitute the bulk of the categorization process. Adopting novel and complex strategies is essential for keeping up with fraudsters and ensuring effective fraud detection.

1.2. RESEARCH CONTRIBUTION

In this light, the goal of this study is to create a reliable prediction model for correctly identifying fraudulent transactions on an imbalanced dataset. We will leverage boosting ensemble techniques, including Adaptive Boosting (Adaboost) and various forms of Gradient Boosting Decision Trees like Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and Categorical Boosting (Catboost), to enhance our predictive models. We will also investigate several pre-processing sampling methods, such as under sampling, oversampling, and hybrid sampling, to solve the issue of class imbalance. These methods will be used to balance the classes in our dataset and improve the model's capacity to handle imbalanced input. Additionally, we will use several assessment measures, such as accuracy, F1-score, precision and recall assessing the performance of our model. These metrics offer information about the model's capability to precisely categorise fraudulent transactions and are frequently utilised when working with imbalanced datasets.

Through this investigation and the use of these techniques, our goal is to identify the assessment statistic that aligns best with our specific objectives and requirements, ultimately enabling us to develop a robust and effective fraud detection prediction model that addresses the challenges associated with imbalanced datasets.

1.3. RESEARCH QUESTIONS

- How do various boosting ensemble methods impact the efficacy of fraudulent credit card transaction detection within an imbalanced dataset?
- What are the effects of different sampling techniques on addressing class imbalance in credit card fraud detection, and how do these techniques contribute to creating a balanced dataset for model development?
- What methods can be employed to enhance the interpretability of boosting ensemble models for credit card fraud detection?
- What strategies can be proposed to minimize classification errors and improve the practical usability of the model for fraud detection in real-world scenarios?

1.4. RESEARCH OBJECTIVE

The aim of this study is to evaluate the effectiveness of a fraud detection model employing the boosting approach as an ensemble method to achieve more accuracy.

- Develop and assess the efficacy of various boosting ensemble methods to improve the detection of fraudulent credit card transactions within an imbalanced dataset.

- Address class imbalance by exploring various sampling techniques to create a balanced dataset for model development.
- Explore methods to enhance the interpretability of the boosting ensemble model, including visualization, predictor importance, and ranking, to provide explanations and identify key factors contributing to fraud detection accuracy.
- Analyse false positives and false negatives to identify patterns, gain insights into misclassification, and propose strategies to reduce errors, thereby enhancing the model's practical usability.

2. LITERATURE REVIEW

2.1. CREDIT CARD FRAUD DETECTION.

Due to the substantial amount of transactions made with credit cards, it is impractical to personally check each transaction for fraud. The impacts of frauds have been reduced due to prevention programmes like the card verification code (CVC), prior authorisation, and card activation processes. Fraud prevention is a proactive system that is set up with the goal of neutralizing every instance of fraud. It is carried out to protect the cardholder's funds and ensure that there is no room for fraud to occur. Examples of these mechanisms include Two-

factor authentication, tokenization, EMV-Chip technology, consumer education and awareness, and collaboration amongst financial institutions, merchants, and law enforcement agencies.

Fraud detection systems takes effect when perpetrators manoeuvre the fraud prevention system and begin a fraudulent transaction. The aim of the fraud detection systems, according to Patel & Singh (2013) is to examine every transaction for the potential to be fraudulent despite the prevention mechanisms and to identify fraudulent ones promptly after the fraudster has started to execute an unauthorised transaction.

Credit card fraud detection refers to the process of identifying and preventing fraudulent activities related to credit card transactions. It involves the use of various techniques, algorithms, and technologies to analyse patterns, behaviours, and transaction data in order to detect and mitigate potential fraudulent transaction. Zhang, et al. (2021) regards a fraud detection model as a transactional behaviour predictive model that uses previous transactional behaviour to predict the validity of present transactional activity.

Credit card fraud detection systems have grown greatly over the years, adopting numerous technology and approaches to keep up with growing fraud strategies. Stakeholders including financial institutions, merchant and law enforcement agencies have been invested in developing nouvelle methods to detect fraud year on year in a bid to keep up with the sophisticated systems and technologies of fraudsters.

The first-ever case of credit card fraud detection can be traced back to the late 1960s and early 1970s, credit card fraud detection was primarily done through manual processes and verification checks. This involved reviewing paper-based transaction records, verifying signatures, and cross-referencing transactions against cardholder information. These manual processes were time-consuming and often prone to errors, making it difficult to detect and prevent fraud. However, they were the only available methods at the time, and financial institutions relied heavily on them to prevent fraud. In the 1980s and 1990s, Rule-based systems were introduced to automate fraud detection processes. Bhattacharya & Sarkar (2018) explained that the rule-based system is a method where fraud is detected using predefined rules and thresholds to flag suspicious transactions based on specific criteria.

The neural networks model became popular in the late 1990s and 2000s, as it paved way for more sophisticated fraud detection capabilities. These models utilized artificial neural networks to learn patterns and detect anomalies in credit card transaction data. As the years go by, there has been tremendous advancement in the use of modern technology to detect fraud, the neural network models were improved upon to ensure accuracy of predictions. Behaviour analysis techniques were incorporated to analyse patterns such as spending habits, transaction frequency and location to identify suspicious activities. According to Phua, et al. (2012), advanced analytics and big data has allowed for more sophisticated models and machine

learning algorithms such as decision trees, random forests, support vector machines and logistic regressions. These algorithms can leverage large amount of data, and utilize different techniques to identify fraud patterns, detect anomalies and suspicious activities in real time.

Recent advancement in fraud detection involves the use of hybrid approaches that combine multiple techniques to enhance fraud detection accuracy and reduce false positives. With the increasing speed and complexity of fraud attacks, real time monitoring systems that analyse transactions as they occur leveraging artificial intelligence (AI) algorithms to detect and respond to fraudulent activities in near real-time have become pertinent. Credit card fraud detection is an ongoing process, and advancements in technologies like artificial intelligence and data analytics continue to shape the landscape of fraud detection.

Credit card fraud detection systems play a crucial role in safeguarding financial transactions and protecting cardholders from fraudulent activities. As stated in an article by FICO, a credit score issuing corporation, fraud detection systems integrate and analyse diverse data sources, including transactional data, user behaviour, and external databases, to identify fraudulent patterns and detect suspicious activities (FICO, no date). Credit card fraud detection system allows for prompt identification and prevention of fraudulent transactions which helps minimize loses for both cardholders and card issuers. These systems have been structured to act as an early warning system by identifying anomalies, enabling timely intervention to prevent fraudulent transactions. This will protect the cardholders and foster customer trust and loyalty; this will also reduce the value of losses in the economy.

While credit card fraud detection systems have their merits, they also face certain challenges and difficulties. Fraudulent transactions are not as prevalent as genuine transactions, thus, resulting in severe class imbalance in the dataset used to run the models for prediction. This imbalance makes it difficult to accurately identify fraudulent transactions and can lead to false positives or false negatives. Studies show that in most fraudulent cases, about 0.17% are fraudulent transactions, making the non-fraudulent class dominant with about 99% (Li, et al., 2022). Striking a balance between fraud protection and consumer privacy can be a difficult issue, as customers are increasingly worried about the security and privacy of their data (Ngai, et al., 2011). This makes it somewhat difficult to use real life data and testing a model on synthetic data may not be as effective or realistic as on real time data. As there are a myriad of locations and e-commerce sites where a credit card may be used, which causes the behaviour to alter, pattern recognition complexity is another difficulty in detecting fraud (Sahin & Duman, 2011). Fraudsters frequently alter their ways to avoid detection, making it difficult for fraud detection systems to stay up with developing fraud patterns. It is important to develop processes and enhance models to facilitate staying ahead of new and sophisticated fraud schemes.

2.2. ADDRESSING IMBALANCED DATASETS.

To guarantee that the model learns well from both fraudulent and non-fraudulent cases and properly predicts rare instances of fraud, the issue of imbalanced datasets must be addressed in credit card fraud detection. Imbalanced dataset occurs when the number of fraudulent transactions is much smaller than the number of genuine transactions, which makes it difficult for fraud detection algorithms to learn and precisely identify fraudulent patterns. A dataset is imbalanced if the classification categories are not almost equally represented. Class imbalance issues have become one of the data mining community's biggest challenges in recent years. This situation is important because it occurs in many real-world classification issues, including fault diagnosis, anomaly detection, medical diagnosis, e-mail foldering, face recognition, and the detection of oil spills, among others (Galar, et al., 2012).

The predictions made by machine learning algorithms when subjected to imbalanced datasets are biased and have misleading accuracy, this is because of lack of information of the minority class. Mishra (2017) explains that the machine learning algorithms assume datasets are balanced with equal class weights, and therefore tends to classify every test case sample into the majority class in order to improve the accuracy metric. Various methodologies and processes have been developed to effectively address class imbalance and achieve quality evaluation of models.

Fernandez, et al. (2018) describes four methods at which class imbalance can be handled namely data level, algorithmic level, cost sensitive learning and ensemble learning. At the data level, balancing approaches are employed as a pre-processing step before any algorithm is done to rebalance the dataset or to reduce the variation between the two categories of data (Dal Pozzolo, et al., 2014). The most common method for dealing with class imbalance is at the data level, which includes several methods for equitably redistributing the classes of data using oversampling, under sampling, and combined sampling.

Oversampling is the process of producing data with a small number of records in a class and balancing it with data from a class with a big number of records to balance the quantity of data records for each class (Chamidah, et al., 2020). Chawla, et al. (2002) experimented the Synthetic Minority Oversampling Technique (SMOTE) on nine different imbalanced datasets to produce minority instances from randomly selected k-nearest neighbours using Naïve Bayes Classifiers. Kotsiantis, et al. (2006) explored the random oversampling method in their study on handling imbalanced datasets, stating that the aim for using this method is to balance class distribution by randomly replicating the minority class. They further emphasized that this could lead to overfitting as it makes the exact copies of the minority samples. He, et al. (2008) presented the Adaptive Synthetic (ADASYN) approach as an extension of SMOTE to adjust

the spread of data and minimise bias by assigning weights to establish a synthetic minority class, this approach was tested on five real-world machine learning datasets, comparing the performance to decision tree models and SMOTE.

Oversampling techniques for balancing datasets is a widely used approach due to its capacity to improve model evaluation. However, this technique has certain limitations, which includes increasing the amount of training samples increases learning time, increases computing costs, and may result in model overfitting (Liang & Zhang, 2012).

On the other hand, the purpose of the undersampling approach is to get a more equal representation of both classes in the dataset by diminishing the dominance of the majority class by randomly picking a subset of examples from the majority class to match the number of instances in the minority class.

Mishra (2017) utilized the random under sampling technique in his study on handling imbalanced dataset on a randomly generated dataset with 50000 samples and 20 variables, he stated that out of all techniques used, the random under sampling technique performed better with the most balanced result and a good trade-off between false positives and false negatives. Tomek Link is another under sampling technique introduced by Ivan Tomek in 1976, it is used to remove the overlap between the classes by removing the majority class sample. (Tomek, 1976). Studies claim that Tomek linkages is not a self-sufficient under sampling technique (Ibrahim, et al., 2021; Zeng, et al., 2016) because it is prone to loosing important information. The Edited Nearest Neighbour (ENN) under sampling method inspects the class labels of a sample's nearest neighbours and removes the misclassified samples from the majority class, these samples are removed iteratively until no further improvements are observed. Wilson (1973) proposes the ENN rule that if the K-Nearest neighbour's samples of a sample differ from their own category, they should be deleted. Under sampling techniques have significant drawbacks despite being timesaving, effective, and generating the best results, including overfitting the model due to the limited amount of data available to train it, the possibility of many majority class samples being omitted, discarded, or eliminated, and the loss of potentially crucial information (Liang & Zhang, 2012).

Combined Sampling is an approach that works by leveraging the strengths of different sampling techniques to achieve better balance between the classes and improve the performance of classifiers on imbalanced datasets. A combination of SMOTE oversampling technique and Tomek Links was proposed by Zeng, et al. (2016) to address the problem of imbalanced dataset, using three medical datasets of common diseases including diabetes, Parkinson's disease and vertebral column. The results shows that the evaluation metric with combined SMOTE-TOMEK was much more improved than a single SMOTE technique. Yang, et al. (2022) also carried out a study on missed abortion datasets to evaluate the results of

combining SMOTE and Edited Nearest Neighbour (ENN) sampling techniques for missed abortion diagnosis. The result shows that SMOTE-ENN has best sampled effect because not only did it synthesize the minority sample, the noisy sample in the majority was also deleted. Zhu, et al. (2020) also combined the ENN undersampling and ADASYN oversampling techniques in their study on lysine succinylation to reduce variance in training sample, and this method gave the most quality evaluation. The combined sampling method is prone to overfitting the model, loss of information and increased computational complexity because of the hybrid nature of the technique.

Handling imbalanced datasets at the algorithmic level entails altering basic algorithms or introducing specialised strategies throughout the training phase to overcome class imbalance. When dealing with decision trees, the adjustment may involve altering the probability estimate at the tree leaf, adjusting the decision threshold, and switching to recognition-based learning rather than discrimination-based learning (Kotsiantis, et al., 2006). In an effort to balance the variation between majority and minority samples, Hartono, et al. (2018) proposed the use of Weighted-SMOTE and Biased Support Vector Machine (BSVM). Although modifying algorithms to address imbalanced data could enhance classification performance, it could contribute complexity to existing algorithms, making them more difficult to implement, tune and interpret (Chawla, et al., 2002).

Cost-sensitive learning involves assigning higher cost to misclassification of the minority class and lower cost to misclassification of majority class in order to reduce bias and balance the importance of different classes. The objective of cost-sensitive learning, according to (Kotsiantis, et al., 2006) is to reduce the cost of misclassification, which may be attained by selecting the class with the lowest conditional risk. Cost-sensitive learning frameworks lay between data and algorithm level techniques, as mentioned by Galar, et al. (2012). It combines algorithmic level adjustments (by altering the learning process to accept fees) and data level transformations (by adding cost to instances). In their investigation of the Hybrid Classifier Ensemble for Imbalanced Data, Li, et al. (2021) makes use of the cost-sensitive classification approach to solve the problem of incompleteness of information by changing the weights of misclassified minority samples rather than the majority ones. By simply modifying the misclassification cost, Fernandez, et al. (2018) propose that, in addition to Support Vector Machines, Artificial Neural Networks and K-Nearest Neighbours are commonly utilised cost-sensitive approaches. The fact that cost sensitive approaches to addressing imbalanced datasets are susceptible to outliers and noise might result in subpar performance on actual data sets.

Ensemble learning is a powerful approach for handling imbalanced datasets. By combining multiple base classifiers or models, ensemble methods can effectively address class imbalance and improve classification performance. According to Priscilla & Prabha (2020),

ensemble learning has consistently proven to be the best method for predicting the class of any practical binary classification issue. In this case, credit card frauds are detected from a significant amount of transactional data because there aren't many fraudulent transactions. The majority of ensemble techniques change the sample data inside bagging and boosting to balance the class distribution of each ensemble member's training set (Li, et al., 2021).

Boosting algorithms are iterative algorithms that give the training distribution a changing weight with each iteration. The weights attached to the improperly categorised instances are increased after each iteration, while the weights attached to the samples that were correctly identified are decreased. It effectively modifies the distribution of the training data; hence it might be regarded as a form of sophisticated sampling approach (Kotsiantis, et al., 2006). Huang, et al. (2005) explored different boosting algorithms with different dimensionalities to provide insight on how these algorithms perform with imbalance datasets.

Bagging, an acronym for Bootstrap aggregating is an ensemble method that handles imbalanced datasets by independently training different models and then aggregate the predictions of individual models through voting in order to achieve the final prediction. It was introduced by Breiman (1996) as a technique used to improve prediction accuracy and model performance. He & Garcia (2009) explored the bagging ensemble method, and its effectiveness when dealing with imbalanced datasets. Zareapoor & Shamsolmoali (2015) also utilized the bagging ensemble method as a technique to detect credit card fraud in an imbalanced dataset, they concluded that the method yielded quality evaluation and takes less computational time.

Recent studies have shown that the best approach for handling imbalanced datasets involves a combination of data-level techniques, algorithmic modifications, cost-sensitive learning, and ensemble learning. This is due to the unique characteristics of pre-processing the data and adapting the algorithms to create an effective model. Galar, et al. (2012) claim that due to the accuracy-oriented nature of ensemble learning methods, when applied directly to imbalanced datasets, they do not resolve the issue that the basic classifier was created to address. Before learning each classifier, it is more useful to pre-process the data with sampling techniques to enhance the sample distribution of the data. Yang, et al. (2022) suggest that combining ensemble techniques with data-level approaches will change the sampling inside bagging and boosting such that the training set's class distribution for every member of the ensemble will be balanced.

The goal of this study is to investigate a unique strategy for addressing imbalanced datasets by fusing ensemble methods and data-level approaches. The suggested method specifically makes use, SMOTE-ENN hybrid sampling, undersampling and oversampling techniques,

accurately identifying credit card fraud on a balanced dataset along with improving the model's assessment metrics.

2.3. TRADITIONAL VS ENSEMBLE MACHINE LEARNING METHODS FOR CREDIT CARD FRAUD DETECTION.

Machine learning has been a hot topic in the field of fraud detection because of its capacity to uncover patterns and correlations, enhance consumer segmentation and targeting, and ultimately raise a business' earnings, development, and position in the marketplace (Deloitte Access Economics, 2017). The use of machine learning has been proven to be an effective strategy for detecting fraudulent conduct. This is obviously feasible since machine learning approaches are data driven. Machine learning may save significant human labour when vast volumes of data are entered into the database. Models are trained on this data and generate the most appropriate output based on the input data (Sharma & Shah, 2021). The progress from using basic rule-based frameworks to more advanced models has enabled the detection of fraudulent activities with higher precision and effectiveness.

Traditional machine learning methods such as rule-based systems rely on pre-defined rules to flag suspicious activity. They have been widely used for solving a wide range of problems, from image recognition to financial predictions. Traditional machine learning algorithms can be categorized into supervised, unsupervised, and reinforcement learning methods (Sharma & Shah, 2021). The supervised approach requires labelled data to train models for classification or regression tasks. Decision tree, Naive Bayes, k-Nearest Neighbour, and Support Vector Machine are some of the most widely used supervised learning techniques that have proven their effectiveness in various fields such as image recognition, text classification and fraud detection (Saranya & Priyadharshini, 2021). On the other hand, unsupervised learning algorithms have been increasingly used in data analysis to identify patterns and relationships between variables. Clustering, Principal Component Analysis (PCA), and Association Rule Mining are among the most employed algorithms for this task. Finally, reinforcement learning is used to teach agents how to behave in an environment based on rewards they receive for certain actions. The algorithm learns to maximise a reward signal by doing acts that result in good outcomes while avoiding actions that result in negative consequences. Q-learning, policy gradient techniques, and actor-critic methods are instances of reinforcement learning algorithms (Sharma & Shah, 2021).

Traditional machine learning methods are computationally efficient and can handle large-scale datasets. Techniques like decision trees, random forests, and linear models have relatively low computational complexity, allowing them to process large amounts of data efficiently (Breiman, 2001). They often produce models that are relatively easy to interpret and understand. This is

particularly important in domains where interpretability is critical, such as healthcare or finance. Decision trees, logistic regression, and support vector machines (SVM) are examples of interpretable models.

While traditional machine learning approach work well in certain areas, they also suffer from weaknesses that limit their effectiveness. For instance, one problem with using single models is their tendency towards overfitting or underfitting data samples due to bias-variance trade-off issues. To address this challenge ensemble techniques can be utilized. These are approaches that merge multiple models to improve accuracy by reducing variance and increasing diversity. Dong, et al. (2020) noted that traditional machine learning approaches may not be effective in dealing with complex data like imbalanced, high-dimensional noisy data; however, their study found that ensembles were able to perform well under such conditions. In fact, combining various models together through the ensemble method has shown time and again its efficacy towards improving performance measures for all types of classification problems.

The application of ensemble machine learning has drastically transformed the approach towards data modelling, prediction, and classification issues. It is an innovative method that integrates several models to enhance the accuracy of predictions while curbing overfitting. The use of ensemble techniques has gained significant traction in diverse fields owing to their proven ability to outperform individual models. A multitude of algorithms have been suggested by researchers for developing ensembles including Bagging (Bootstrap Aggregating), Boosting, Stacking (Stacked Generalization) among others which demonstrate improved accuracy when tested against traditional machine learning methods. In addition to achieving better performance in predictive tasks; Ensemble also addresses common issues like overfitting, bias-variance trade-off with classical machine learning approaches thereby making it attractive for practical applications where there are limited datasets or complex feature spaces.

Mishra & Ghorpade (2018) argue that ensemble machine learning methods are superior in detecting credit card fraud because they can adapt to new patterns of fraud in real-time. This is a crucial feature because it enables financial institutions to detect and prevent fraudulent transactions before significant damage occurs. Furthermore, ensemble machine learning methods have proven successful in identifying previously undetected types of credit card scams such as account takeover attacks or phishing scams. According to Dong, et al. (2020), using an ensemble model can also reduce false positives, resulting in fewer incorrect alerts being raised for human review. These benefits make ensemble machine learning particularly appealing in fraud detection applications both from a performance perspective and given its potential impact on business costs related to fraudulent activity. Therefore, incorporating these new techniques into existing systems could lead not only to improved security but also increased efficiency which can be beneficial across many industries including banking, e-commerce, retail among others.

The use of ensemble machine learning methods has become increasingly common due to their improved performance compared to traditional algorithms. However, this approach requires substantial amounts of training data and expertise in algorithm selection and parameter tuning. It will also require more computational resources and time as it deals with building two or model at the same instance.

While both approaches have their pros and cons, it is evident that ensemble machine learning methods have proven to be more effective in detecting fraudulent transactions even though they require more expertise and resources. Credit card fraud remains prevalent globally, despite the numerous advancements made in this field over time. Hence the need to continue exploring new techniques to enhance its prevention.

2.4. BOOSTING ENSEMBLE MACHINE LEARNING

The Boosting Ensemble Machine Learning (BEML) approach has developed as a useful tool for detecting fraudulent transactions with high accuracy rates in recent years. BEML combines different algorithms to maximise their strengths and improve the model's overall performance. Research studies suggest that BEML outperforms traditional machine learning methods such as Support Vector Machines (SVM) and Random Forests in this area. Therefore, understanding how these models work and their relevance to fraud detection is essential in curbing financial crimes effectively. Boosting ensemble machine learning methods represent a promising approach to improving fraud detection accuracy. This method involves combining multiple weak classifiers to create a strong classifier that can better identify patterns of fraudulent behaviour.

A study on boosting learning by Sun, et al. (2006) provides an excellent example of how boosting can be applied effectively to detect multiple types of fraudulent activities with varying degrees of severity. By leveraging the strengths of individual classifiers while mitigating their weaknesses, boosting improves overall prediction accuracy by reducing bias and variance in the model. It achieves this by iteratively adjusting weights assigned to each classifier based on its performance until it reaches optimal predictive power. The resulting strong classifier is more robust than any individual component model alone, making it particularly effective at identifying complex patterns indicative of fraudulent activity.

As Willis (2020) points out, digital fraudsters often use sophisticated techniques to evade detection by traditional rule-based systems. Boosting algorithms offer an alternative solution because they learn from data sets rather than relying on pre-established rules. Thus, they can adapt quickly to new threats and identify previously unknown forms of deception. Researchers today have employed various types of boosting techniques including AdaBoost proposed by

Freund & Schapire (1997), Gradient Tree Boosting introduced by Friedman, et al. (2000), XGBoost developed by Chen & Guestrin (2016), LGBM developed by Ke, et al. (2017) and Catboost introduced by Prokhorenkova, et al. (2018).

Adaboost

AdaBoost, short for Adaptive Boosting, is an effective Ensemble machine learning technique that combines multiple weak learners to create a strong classifier, has become increasingly popular in recent years due to its high accuracy and versatility (Freund & Schapire, 1997). The concept of ensemble learning is based on the idea of integrating data fusion, modelling and mining into a unified framework that aims to improve the performance of prediction models. This technique is particularly effective in solving binary classification problems, where the goal is to assign each input instance to one of two classes like the credit card fraud detection instance because of its exceptional accuracy achieved through iterative optimization processes (Freund & Schapire, 1997).

According to Dong, et al. (2020), the Adaboost technique involves extracting features using various transformations followed by utilizing several learning algorithms that produce weak predictive results. These results are then fused together using voting schemes in an adaptive way, resulting in better predictive performance and knowledge discovery. The Adaboost algorithm works by combining decision trees or other base classifiers with different weights assigned based on their individual performance. As each subsequent tree is built and evaluated, it focuses more on misclassified samples from previous iterations until no further improvement can be made.

Adaboost has been proven to be a versatile and efficient ensemble method for classification problems in machine learning. Li, et al. (2018) reported that Adaboost has successfully been applied in various domains, including computer vision, natural language processing, and bioinformatics. The algorithm works by iteratively training weak classifiers on weighted datasets and subsequently combining their predictions through a weighted sum to form the final prediction model. This process provides high accuracy while minimizing overfitting of the data.

As pointed out by Zhu, et al. (2006), Adaboost technique can efficiently identify the most important features for classification purposes. However, there are situations when working with imbalanced datasets can lead to subpar results. That's where resampling and ensemble techniques come into play. According to Snieder, et al. (2021) recent paper, utilising these approaches has demonstrated the possibility of improvements in the accuracy of models of imbalanced datasets like streamflow data utilised in hydrological prediction. In fact, their research discovered that the use of ensemble techniques would be beneficial for minimising both amplitude and timing errors in extremely imbalanced flow datasets. This suggests that

incorporating ensemble techniques like Adaboost could be a valuable tool not only for selecting features but also improving overall model performance on challenging data sets such as those seen in hydrology and other scientific fields dealing with highly imbalanced data distributions.

Adaboost's success across different fields highlights its effectiveness as an ensemble method for classification tasks in machine learning. Its ability to handle noisy and imbalanced data with simplicity makes it an ideal choice for many applications requiring high accuracy without sacrificing computational efficiency.

Gradient Boosting

Gradient Boosting Decision Trees (GBDT) often referred to as Gradient Boosted Trees or simply Gradient Boosting, is a widely used technique in the field of machine learning that has been proven to be effective for both classification and regression problems. It works by combining multiple decision trees to create a more accurate and predictive model. Jerome H. Friedman outlines the basic principles of this approach and highlights its impressive capabilities (Friedman, 2001). This algorithm is notable for its ability to combine multiple weak models into a single strong model that can make highly accurate predictions about complex datasets. However, it's crucial to note that the success of any gradient boosting machine depends heavily on the quality and suitability of the individual models being combined.

According to Jablonka, et al. (2020), "GBDT combines multiple decision trees to create a more accurate and predictive model." This means that it uses an iterative approach where each new tree corrects errors made by previously trained models until no further improvements are possible. The high degree of accuracy achieved with GBDT can be attributed to its ability to handle complex data sets with both numerical and categorical features, as well as being robust against overfitting. Furthermore, the interpretability of the model allows users to understand how each feature contributes towards the final prediction, making it easier for them to identify important variables when making decisions based on this data.

GBDT has proven effective for handling imbalanced datasets like in credit card data. Chen & Guestrin (2016) stated that the algorithm assigns higher weights to minority class samples, which allows it to effectively identify fraudulent activity with high precision, by doing so, the algorithm ensures that no suspicious transaction goes unnoticed and helps prevent significant financial losses for both customers and credit card companies alike. Furthermore, the ability of gradient boosting decision tree to handle imbalanced datasets also makes it a valuable tool in other areas where such data imbalance is common, including medical diagnosis and customer churn prediction.

According to Ye, et al. (2009), GBDT offers interpretability that enables analysts to understand the underlying patterns and features contributing to fraudulent transactions better. This insight

can facilitate developing more effective fraud prevention strategies by identifying new variables or adjusting existing algorithms based on the learned insights from interpretable models like GBDTs. Furthermore, interpretability allows analysts to identify false positives and negatives when dealing with fraudulent transactions effectively.

XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful and efficient implementation of the gradient boosting algorithm. It was developed by Tianqi Chen and his team at the University of Washington and was first introduced in 2014 (Chen & Guestrin, 2016). XGBoost is designed to deliver high performance and handle large-scale datasets efficiently. It has been gaining popularity due to its ability to optimize model performance and improve accuracy. Its strength lies in the combination of gradient boosting and regularized regression techniques, which work together to minimize prediction errors by iteratively adjusting the weights assigned to each feature.

The versatility, ease of use, and robustness of this technique makes it an ideal candidate for a wide range of applications. Whether it's image classification, natural language processing, or recommendation systems - XGBoost can handle them all with great efficiency (Chen & Guestrin, 2016). A recent study by Zhao, et al. (2020) where the authors evaluated six different models to predict multiple sclerosis disease, demonstrated that XGBoost, along with LGBM, were superior in terms of performance. However, the effectiveness of this approach can vary depending on the context and dataset used for training. It is important to note that while XGBoost offers impressive accuracy results across a wide range of applications, it requires careful tuning of hyperparameters and preprocessing steps for optimal performance.

This is where resampling and ensemble techniques come into play, as they have shown to significantly improve model performance on imbalanced datasets (Snieder, et al., 2021). By combining multiple models trained on randomly sampled subsets of the data or using different algorithms altogether and averaging their predictions, ensemble methods can effectively reduce overfitting while improving generalization capabilities. Similarly, resampling approaches such as oversampling minority classes or undersampling majority classes can help balance dataset distributions and prevent bias towards dominant categories. Thus, while XGBoost's optimization techniques are undoubtedly beneficial for many applications, it's essential not to overlook complementary strategies like ensemble learning when working with complex or imbalanced datasets like credit card fraud detection.

LGBM

Light Gradient Boosting Machine, or LGBM, is a cutting-edge gradient boosting algorithm that leverages tree-based learning. Introduced by Ke, et al. (2017) as an open-source initiative,

LGBM has significantly transformed the landscape of machine learning. It operates by training a series of weak models, called decision trees, in sequence. This approach not only enhances model performance but also effectively addresses the limitations present in earlier gradient boosting libraries like XGBoost and GBM.

One of LGBM's standout features is its remarkable efficiency in handling substantial datasets. This prowess is attributed to its innovative histogram-based tree-building mechanism, which sets it apart. This capability to efficiently process large datasets renders LGBM exceptionally suitable for complex tasks, such as credit card fraud detection. Its efficiency ensures speedy transaction processing, facilitating real-time detection of fraudulent activities.

Another notable advantage of LGBM is its innate ability to handle categorical features seamlessly. Unlike some other gradient boosting algorithms, LGBM eliminates the need for extensive data pre-processing, such as one-hot encoding. It directly manages categorical features, streamlining the pre-processing pipeline and conserving computational resources. Additionally, LGBM employs a novel technique known as Gradient-based One-side Sampling (GOSS) to enhance training speed without compromising accuracy. GOSS intelligently samples instances based on gradient information, prioritizing those with larger gradients. This strategic approach mitigates overfitting and enhances the model's ability to generalize.

However, like all machine learning algorithms, LGBM does come with its own set of limitations. Notably, it exhibits sensitivity to noisy and imbalanced datasets. To circumvent this challenge, proper pre-processing techniques, such as noise removal and data balancing, should be applied.

Researchers like Malik, et al. (2022) have harnessed LGBM's potential by combining it with pre-processing techniques like under sampling and SMOTE to effectively handle imbalanced datasets. This approach has yielded high accuracy and low false-positive rates, demonstrating LGBM's capability in addressing class imbalance. Furthermore, Zhang, et al. (2020) proposed an innovative approach for credit card fraud detection using LGBM on two datasets. This approach was compared with traditional methods like Logistic Regression, SVM, and XGBoost. LGBM demonstrated its prowess by achieving improved accuracy and AUC-ROC scores, underscoring its effectiveness in managing large-scale and complex datasets.

LGBM has emerged as a game-changer in the realm of machine learning, offering speed, efficiency, and versatility. Its ability to handle large datasets, work seamlessly with categorical features, and incorporate advanced techniques like GOSS positions it as a robust tool for tackling real-world challenges, including fraud detection and beyond.

Catboost

CatBoost stands for Categorical Boosting, it is a gradient boosting library specifically designed to handle categorical features effectively. It was developed by the team at Yandex, a Russian multinational technology company (Prokhorenkova, et al., 2018). Catboost uses gradient boosting on decision trees and groups observations by categories which allows it to achieve high accuracy while being resistant to overfitting (Prokhorenkova, et al., 2018).

One area of application that has seen great success using Catboost is the spatial data domain, where it outperforms many other algorithms in terms of accuracy and efficiency (Takacs, 2018). Takacs (2018) argues that this algorithm can be used as an effective tool for assessing the quality of geo-spatial datasets and offers significant improvements over existing methods for data analysis. The reason behind its success lies in its unique approach towards handling categorical variables by converting them into numerical representations through a series of decision trees, which results in improved model performance even when dealing with high-dimensional datasets (Prokhorenkova, et al., 2018).

Alsenaani (2022) have developed a model using these techniques that detects fraudulent transactions and flags them automatically. This approach is useful because traditional methods of fraud detection are often unable to detect complex patterns or anomalies within large datasets without significant human input. Using this type of method can help businesses minimize financial losses due to fraud while also streamlining the detection process overall. The combination of Catboost with other Machine Learning techniques like the ones used by Alsenaani (2022) can lead to even more impressive results when tackling real-world problems where accuracy is crucial. Additionally, new applications of these technologies continue emerging constantly, opening up possibilities for their implementation across diverse industries ranging from finance to healthcare.

Catboost stands out among other machine learning algorithms because it enables efficient handling of categorical features while achieving state-of-the-art performance on numerous real-world problems such as analysing geo-spatial datasets (Takacs, 2018). However, despite its many advantages over other machine learning methods, Catboost still requires careful consideration and interpretation of results. As with any algorithmic model, there are limitations to what it can accomplish depending on the nature of the dataset and specific problem being addressed.

Overall, the combination of weighted ensemble learning, and iterative training make boosting algorithms effective tools for detecting fraudulent activities in the digital age. Their ability to enhance the performance of weaker classifiers ensures higher levels of accuracy while being adaptable enough to detect ever-evolving types of cybercrime.

According to Cao, et al. (2018), boosting is a powerful and efficient approach for detecting fraudulent activities. The results of their study demonstrate its efficacy in identifying various

types of fraud cases. However, as with any algorithmic solution, there are trade-offs that one should consider before selecting this method over other alternatives. Specifically, the authors note that accuracy versus efficiency must be considered when making such a decision. While boosting may provide high levels of accuracy in identifying fraudulent transactions quickly and effectively in real-time environments, it may also come at the cost of increased computational complexity and time consumption (Cao, et al., 2018). Therefore, while boosting algorithms can enhance classification performance on imbalanced datasets significantly; other techniques must also be considered for an effective solution when dealing with such scenarios.

2.5. PERFORMANCE EVALUATION FOR CCFD

Performance evaluation is an important phase in the machine learning workflow. It enables the comparison of multiple methods, the selection of the best-performing model, the discovery of areas for development, and hyperparameter optimisation. It guarantees that the algorithm chosen is appropriate for the job and produces trustworthy results in real situations. Metrics like precision, recall and F1-score are more useful in assessing how well a model performs in identifying fraudulent cases without falsely flagging non-fraudulent ones.

When it comes to credit card fraud detection, precision and recall are two important metrics that help evaluate the performance of a model. However, relying solely on these metrics can lead to biased results because they do not consider false negatives or false positives. This is where F1 score comes in handy as it provides a balanced measure of both precision and recall (Shakya, 2018). The F1 score considers both true positives and true negatives while also considering how many predicted positive cases were actually negative. In other words, the F1 score measures the harmonic mean of precision and recall by using their weighted average. By doing so, it becomes easier to identify fraudulent activities with greater accuracy while minimizing errors such as declining legitimate transactions or approving fraudulent ones. As Han, et al. (2020) suggest, F1 score can be utilized to compare the efficiency of various credit card fraud detection algorithms and models. F1 score provides an overall assessment of precision and recall rates simultaneously since it considers both false positives and false negatives. A higher F1 score indicates better model performance in identifying fraudulent activities while minimizing the number of legitimate transactions mistakenly flagged as fraudulent ones or vice versa (Han, et al., 2020). The use of F1 score enables decision-makers in financial institutions to select a suitable algorithm or model that meets their preferences based on accuracy levels or other specific requirements.

According to Strelcenia & Prakoonwit (2023), "precision is a critical metric for evaluating credit card fraud detection using ensemble methods on imbalanced datasets." They have shown that data balancing techniques, such as B-SMOTE, K-CGAN, and SMOTE can significantly

enhance precision and recall in credit card fraud detection. As pointed out by Rubaidi, et al. (2022), precision has been identified as an effective means for evaluating fraud detection models, given that it measures how often the model correctly identifies actual fraudulent transactions from all flagged transactions. By focusing on precision, fraud analysts can gauge whether their systems generate few false alerts or if they lack sufficient sensitivity and flag too many legitimate cases wrongly. This approach ensures that only potentially fraudulent activities are investigated while minimizing operational costs associated with investigating non-fraudulent activity. However, leveraging precision effectively requires robust data analysis capabilities and sophisticated algorithms capable of detecting complex patterns within large datasets.

Recall measures how well a classifier can identify all instances of a given target class within the dataset. It considers both false negatives and true positives, making it more suitable for imbalanced data where correctly identifying rare events is crucial but challenging. In contrast, precision focuses solely on true positives and can lead to misleading results due to its inability to detect false negatives. Therefore, in order to ensure accurate evaluation of classification models on imbalanced data sets, it is recommended that recall be used as a more reliable metric than traditional methods like accuracy or precision (Rubaidi, et al., 2022). By doing so we can accurately evaluate our models and make informed decisions about their performance which can then be further improved upon through modifications or additional training iterations if necessary.

The confusion matrix is another useful tool that provides insights into how well classifiers perform across different classes during model evaluation (AlEmad, 2022). It helps identify where misclassification occurs and allows for targeted adjustments to improve detection of fraudulent activities. By combining these two powerful techniques, researchers have achieved impressive results in detecting fraud with imbalanced datasets while minimizing false positives – a common issue when dealing with highly skewed data distribution. Rubaidi, et al. (2022) suggest that credit card companies can use precision and recall metrics obtained from a confusion matrix analysis to adjust in real-time based on current trends in fraudulent activity. Furthermore, by refining their fraud detection strategies through careful analysis of the confusion matrix results, credit card companies could improve not only efficiency but also customer satisfaction rates.

According to Penmetsa & Mohammed (2021), the confusion matrix is particularly helpful because it allows for the identification of both false positives and false negatives in credit card fraud detection models.

False positives occur when a transaction is flagged as fraudulent when it isn't fraudulent, while false negatives happen when a fraudulent transaction goes undetected by the model. These

errors can have significant financial consequences for both individuals and businesses alike. To improve the accuracy of credit card fraud detection models, it's crucial to understand where these errors are occurring. By using the confusion matrix, we can see how many true positive predictions (i.e., correctly identified cases of fraud) there were versus false positives (transactions incorrectly identified as fraudulent). We can also see how many true negative predictions there were versus false negatives (fraudulent transactions that went undetected). Armed with this information, we can further optimize our credit card fraud detection models to minimize errors and increase their effectiveness.

2.6. SUMMARY OF LITERATURE

Fraud prevention measures like two-factor authentication, tokenization, and EMV chip tech are in place to safeguard cardholders' funds, while Fraud Detection systems employ predictive models to promptly flag suspicious transactions. These systems have evolved with advanced technologies, including neural networks and machine learning, enabling real-time fraud pattern detection. Combining techniques like ensemble learning enhances accuracy, and AI-driven real-time monitoring is pivotal in fraud detection. Ultimately, these systems bolster trust, loyalty, and financial security for cardholders and issuers.

Imbalanced datasets, where fraudulent transactions are scarce compared to genuine ones, challenge fraud detection algorithms, leading to biased predictions and misleading accuracy. Methods to tackle this include data-level techniques (oversampling, undersampling, combined sampling), algorithm tweaks, cost-sensitive learning, and ensemble methods. Combining oversampling and undersampling can enhance balance but risks overfitting and complexity. Algorithm adjustments and cost-sensitive learning help but might complicate interpretation. Ensemble methods like bagging and boosting are potent for addressing class imbalance and boosting classification performance.

Ensemble machine learning methods, including Adaboost, Gradient Tree Boosting, XGBoost, LGBM and Catboost, enhance fraud detection accuracy by combining models to reduce variance and increase diversity. They outperform traditional approaches and adapt to evolving fraud patterns in real-time, reducing false positives and addressing issues like overfitting and bias-variance trade-offs.

Evaluating credit card fraud detection involves using metrics such as precision, recall, F1 score, and the confusion matrix. These techniques help select suitable algorithms, enhance detection strategies, and improve fraud prevention in transactions.

2.7. GAPS IN EXISTING RESEARCH

The detection of fraudulent transactions in large credit card dataset has been a subject of considerable interest over the years. Numerous studies have explored various machine learning algorithms and techniques particularly ensemble methods to achieve accurate predictions, but there remain some gaps in this existing literature that suggests opportunities for further investigation.

Mishra & Ghorpade (2018) studied fraud detection using ensemble methods (like Gradient Boosted Trees and Random Forests) and classification techniques (like logistic regression, SVM, and decision trees) on a skewed credit card dataset from September 2013 in Europe. To balance the data, they used random undersampling mthod. Their most effective approach was Gradient Boosted Trees, which achieved the highest precision and recall. However, they didn't explore other sampling methods, such as oversampling or hybrid sampling, to further enhance GBDT predictions.

Taha & Malebery (2020) employed a 5-fold cross-validation technique to tackle dataset imbalances in their study on optimizing Light Gradient Boosting Machine for credit card fraud detection. Their findings indicated that the OLGBM approach surpassed other machine learning algorithms in performance. However, their research fell short in providing thorough interpretation and explanation of the key predictors contributing to the model's predictions.

Hancock & Koshgoftaar (2021) investigated Medicare fraud detection using Gradient Boosting Decision Tree algorithms such as Catboost, LGBM, and XGBoost, analyzing two distinct Medicare insurance claim datasets. Their findings revealed Catboost as the most effective performer. However, one limitation of their study was the absence of a combination of GBDT algorithms with sampling techniques; instead, these techniques were paired with classifiers.

Ramani, et al. (2022) examined LGBM and Catboost as Gradient Boosting Decision Tree (GBDT) techniques and assessed their performance against established methods like Auto Encoder, Neural Networks, Logistic Regression, and K-means clustering using a credit card dataset containing 30,000 instances. Notably, LGBM outshined the other models, achieving an impressive accuracy rate of 97%. However, one limitation of their research was the absence of exploration into hybrid models that could potentially further enhance accuracy.

In this research, our primary goal is to bridge existing gaps by tackling the challenge of imbalanced datasets through the implementation of three distinct sampling methods: random undersampling, SMOTE oversampling, and SMOTE-ENN. To comprehensively evaluate and compare our approach, we will also incorporate the Adaboost method alongside a suite of Gradient Boosting Decision Tree algorithms, including GBM, XGBM, LGBM, and Catboost. Our evaluation will be multifaceted, encompassing essential metrics such as accuracy, F1 score, precision, and recall. Additionally, we will utilize the power of the confusion matrix to

dissect false negatives and false positives, offering insights into the model's performance. To gain a deeper understanding of our results, we will delve into predictor importance and other key features. This comprehensive approach will provide a well-rounded assessment of our methods, ultimately contributing to a more robust and informed analysis of imbalanced dataset handling and predictive model performance.

3. RESEARCH METHODOLOGY

This study adopts a quantitative approach for conducting predictive analysis to discern the legitimacy of credit card transactions, distinguishing between fraudulent and non-fraudulent ones. According to Waller & Fawcett (2013), predictive analysis involves utilizing quantitative and computational techniques to create models that anticipate future occurrences, based on historical data analysis. Despite using a synthetic dataset, the dataset is meticulously generated to precisely emulate the patterns and features present in actual data, ensuring the protection of cardholders' personal information. This information will facilitate the execution of predictive analysis.

The purpose of this investigation is to assess the efficiency of a fraud detection model by employing the boosting technique as an ensemble method to achieve enhanced accuracy. The research aims to compare the accuracy and comprehensibility of the model to make it more accessible for stakeholders.

To undertake this investigation, we opted for an experimental research design as our preferred approach. In this design, we conduct controlled experiments to assess the efficacy of various methods, algorithms, or techniques in detecting and preventing fraudulent credit card transactions. The central objective of this design is to pinpoint the most effective and accurate approach for distinguishing between genuine and fraudulent transactions.

The research process is as seen below.

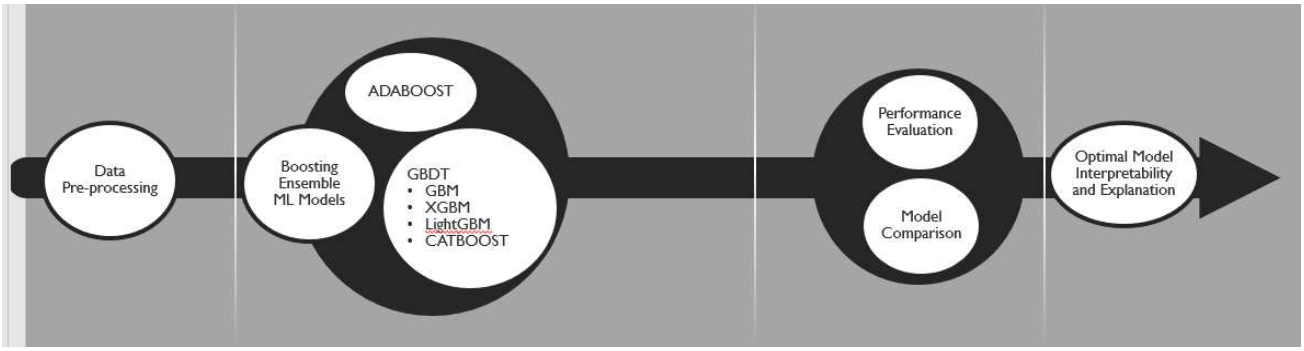

Figure 3: Research methodology

3.1. ABOUT THE DATASET

The data collection process for this research was conducted using the secondary method. A synthetic dataset comprising credit card transactions was obtained, designed to closely resemble actual transactions from financial institutions. This dataset was used in a study conducted by Emily Smith which was obtained from Kaggle. The following key points explains the characteristics of the dataset.

- It consists of purchases made with 693 different businesses and 320 different clients.
- The total number of transactions in the dataset is 975,036.
- There are 5,412 fraudulent transactions and 969,624 non-fraudulent transactions.
- The dataset contains 22 variables with numerical and categorical features.
- Fraudulent transactions are represented as '1' and non-fraudulent transactions as '0'.
- The transactions in the dataset span from February 2021 to March 2022.

```
data.nunique()
```

```
trans_date_trans_time    437161
credit_card_number          320
merchant                    693
category                     14
amount                    47121
first_name                  346
last_name                   478
gender                        2
street                      960
city                        876
state                        51
zip_code                    947
latitude                    945
longitude                   946
city_population             863
job                         491
day_of_birth                945
trans_number             975036
unix_time                958061
merchant_lat             947225
merchant_long            963180
fraud                         2
dtype: int64
```

Figure 4: Characteristics of dataset

3.2. DATA PREPROCESSING

Data pre-processing plays a crucial role in the success of machine learning models, especially when dealing with complex tasks like credit card fraud detection. By employing appropriate data pre-processing methods, we can enhance the performance and interpretability of the ensemble models, ultimately leading to more effective fraud detection.

Categorical variables, such as category or transaction types, require special treatment before feeding them into the boosting ensemble model. Techniques like label encoding or one-hot encoding can be applied to transform categorical variables into numerical representations, allowing the model to process the information effectively.

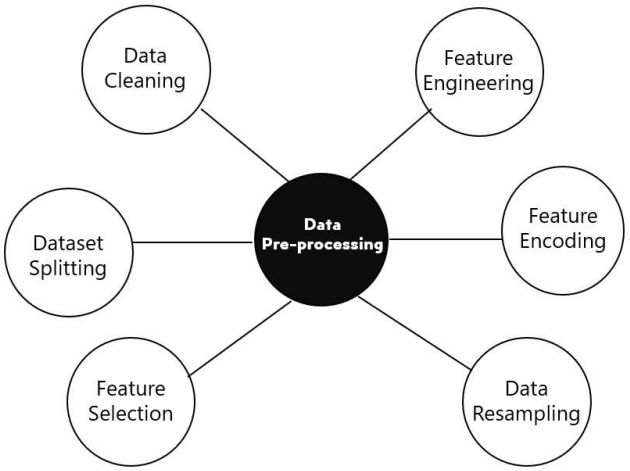

Figure 5: Data pre-processing techniques

3.2.1. Data Cleaning

The dataset contains 22 columns of both categorical and numerical values with 975,036 transactions. Generally, datasets often contain missing values, which can adversely impact model performance. Identifying and removing outliers is essential to maintain model stability. This synthesized dataset doesn't contain any null or missing values.

```
In [10]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 975036 entries, 0 to 975035
Data columns (total 22 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   trans_date_trans_time  975036 non-null  object
 1   credit_card_number     975036 non-null  float64
 2   merchant               975036 non-null  object
 3   category               975036 non-null  object
 4   amount                 975036 non-null  float64
 5   first_name             975036 non-null  object
 6   last_name              975036 non-null  object
 7   gender                 975036 non-null  object
 8   street                 975036 non-null  object
 9   city                   975036 non-null  object
 10  state                  975036 non-null  object
 11  zip_code               975036 non-null  int64
 12  latitude               975036 non-null  float64
 13  longitude              975036 non-null  float64
 14  city_population        975036 non-null  int64
 15  job                    975036 non-null  object
 16  day_of_birth           975036 non-null  object
 17  trans_number           975036 non-null  object
 18  unix_time              975036 non-null  int64
 19  merchant_lat           975036 non-null  float64
 20  merchant_long          975036 non-null  float64
 21  fraud                  975036 non-null  int64
dtypes: float64(6), int64(4), object(12)
memory usage: 163.7+ MB
```

Figure 6: Missing values in dataset

3.2.2. Data Splitting

In the machine learning workflow, splitting the data before pre-processing is a crucial and foundational step with several key benefits. Firstly, it helps prevent data leakage, which occurs when information from the test set unintentionally influences pre-processing steps, leading to biased models and overestimation of performance. By keeping the test set isolated, we can simulate real-world scenarios where the model faces unseen data, enabling us to assess its generalization capabilities accurately.

Furthermore, the split allows us to objectively evaluate and compare different models on an independent dataset, aiding in the selection of the best-performing model for deployment. It also helps in identifying and addressing overfitting issues, as we can assess a model's performance on data it has not seen during training, providing insights into its ability to generalize beyond the training set.

In the process of developing the model, performing cross-validation on the training set becomes crucial as it facilitates hyperparameter tuning without affecting the integrity of the test set. This allows for fine-tuning the model and optimizing its performance effectively. The dataset was split into an 80% training set and a 20% testing set, ensuring a substantial amount of data is available for the model to learn meaningful patterns and relationships from the input features.

```
from sklearn.model_selection import train_test_split

# Separate the features and target variable
X = data.drop(['fraud'], axis=1)  # Features (all columns except 'fraud')
y = data['fraud']  # Target variable

print("Dataset Shape:", X.shape, y.shape)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=42)

print("Train and Test Shape:", X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
Dataset Shape: (975036, 21) (975036,)
Train and Test Shape: (780028, 21) (195008, 21) (780028,) (195008,)
```

Figure 7: Splitting the dataset.

3.2.3. Dealing with Outliers

Kwak & Kim (2017) define Outliers as the datapoints lying far away from the majority of other data points. Outliers can have a significant impact on statistical analysis and machine learning models, they can skew the distribution, affect parameter estimation, and introduce bias in predictions. Identifying and handling outliers is essential to ensure the accuracy and reliability of detecting credit card fraud with the dataset.

In the graph representing transaction amounts, most transactions fall within the range of 0 to around 2500, forming a cluster of data points. However, there are a few transactions with unusually large amounts that lie far away from the majority of transactions.

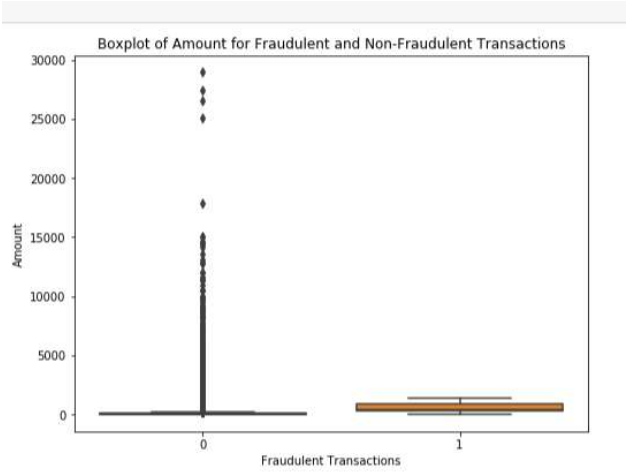

Figure 8: Outliers in dataset

When dealing with outliers, particularly if they represent a minor fraction of the data, excluding them from our analysis can be a justifiable approach. Nevertheless, it is essential to proceed with caution and ensure that these outliers do not signify fraudulent transactions. By employing the interquartile range method, we have identified more than 100,000 transactions flagged as outliers, of which approximately 76% are fraudulent transactions removed from the dataset.

While excluding outliers may help improve the generalization of our analysis, eliminating a large number of fraudulent transactions raises concerns about the accuracy and reliability of our fraud detection model. Striking a balance between outlier removal and preserving crucial fraudulent data is imperative to maintain the effectiveness and credibility of our predictive analysis. Additionally, considering alternative techniques and robust algorithms that can handle outliers effectively could enhance our ability to identify and address fraud without compromising the integrity of our dataset.

3.2.4. Feature Engineering

The 'day_of_birth' column is converted to age because age can be a more meaningful and informative feature than the exact birthdate. This will create a more informative and suitable feature for fraud detection tasks, making it easier for machine learning models to capture relevant patterns and relationships between age and fraudulent transactions.

```
# Convert 'day_of_birth' to a datetime object
data['day_of_birth'] = pd.to_datetime(data['day_of_birth'])

# Calculate the current year
current_year = datetime.now().year

# Calculate age by subtracting the birth year from the current year
data['Cardholder_age'] = current_year - data['day_of_birth'].dt.year

data.head()
```

latitude	longitude	city_population	job	trans_number	unix_time	merchant_lat	merchant_long	fraud	cardholder_age
41.4802	-86.6919	1423	Psychologist, forensic	0442179ed2d64942bc45c5442c3d5895	1328054544	41.587290	-87.582956	0	25
43.0172	-111.0292	471	Education officer, museum	ef5b8cbad64ceecb8d1b81960a49b467	1328054607	42.032369	-111.700448	0	56
36.6966	-96.7869	471	Sub	eae42972de9670a3d907035c0594d494	1328054744	36.413203	-95.986495	0	82
40.8027	-81.3739	192805	Building control surveyor	2832eb7d75a6f04468c3e52b03103090	1328054781	40.576060	-81.529572	0	60
35.9866	-106.0654	18408	Historic buildings inspector/conservation officer	345575e8ebaaaed20c5ded6694383a3a	1328054900	35.125220	-105.981956	0	51

Figure 9: Feature engineering for day of birth

Understanding the correlation between the age of cardholders and fraudulent transactions is straightforward. The visual representation reveals that most fraud victims fall within the age range of late 20s to late 60s, commonly referred to as the middle-aged working class. This demographic is characterized by financial independence and sufficient earnings, which may lead to impulsive shopping behaviour. Consequently, they become more vulnerable to scams, as the fraudulent amounts might not be substantial relative to their account balances.

```
: data[data.fraud==1]['Cardholder_age'].hist(bins=10)

: <matplotlib.axes._subplots.AxesSubplot at 0x2402577d5c8>
```

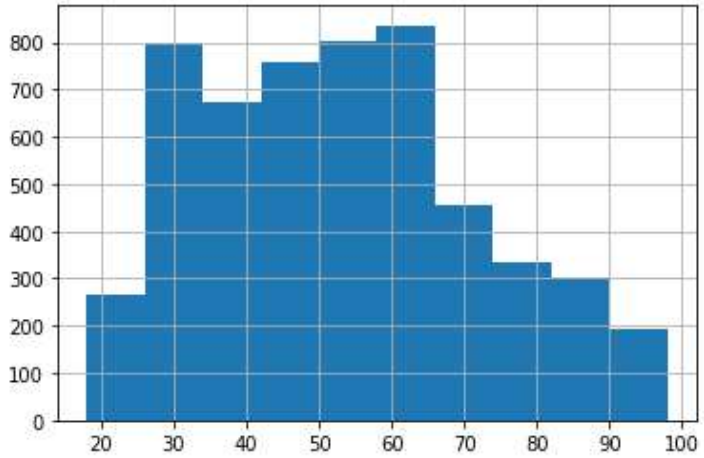

Figure 10: Histogram showing cardholders' age for fraudulent transactions.

For the date of each transaction, feature engineering is done to extract the hour, day, month and year in order to capture temporal patterns or trends that might exist in the dataset. The extracted date and time features can be to detect unusual or irregular patterns in transaction timing could potentially indicate fraudulent behaviour.

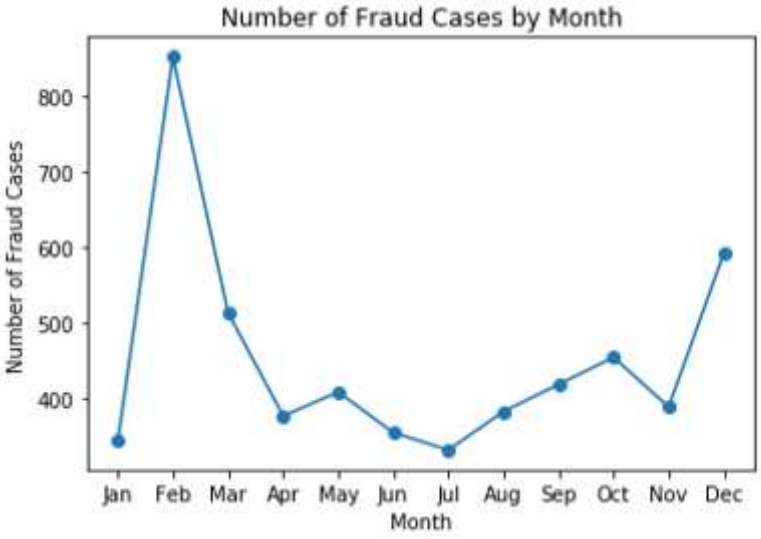

Figure 11: Graph showing number of fraud cases per month.

The analysis reveals that January has the highest number of fraud cases. This observation aligns with the post-holiday season, where there is typically a surge in both online and in-store transactions. During this period, fraudsters may capitalize on the increased volume of transactions to target unsuspecting consumers and merchants. Furthermore, after the holiday season, there is often a rise in product returns and chargebacks. Fraudsters can exploit this

process by engaging in fraudulent activities, such as returning stolen or counterfeit goods to obtain refunds. The financial strain experienced by many individuals after the holiday season, due to heightened spending, may also contribute to an increase in fraudulent activities, including credit card fraud, as some people resort to deceptive means to manage their financial obligations.

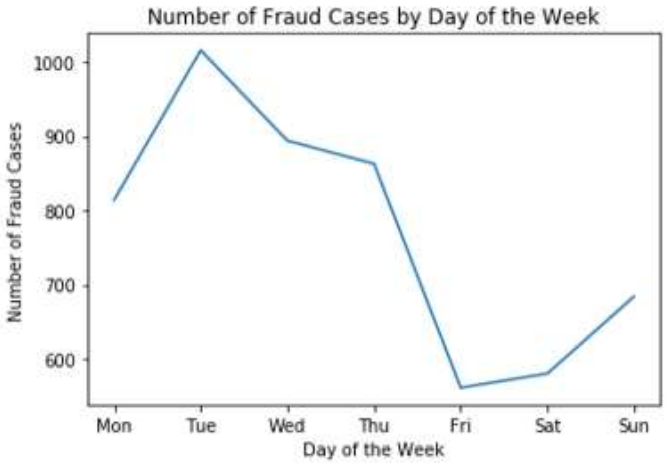

Figure 12: Graph showing number of fraudulent cases weekly.

The above shows that Tuesday is the day of the week that records the highest amount of fraudulent cases. The reason could be due to historical transactions that has been successful on that day.

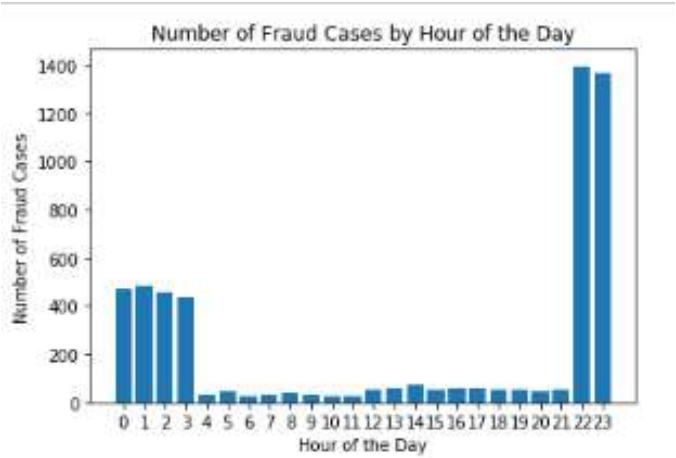

Figure 13: Graph showing number of fraudulent cases hourly.

The higher prevalence of fraud cases between 10 and 11pm can be attributed to the exploitation of time zone differences by fraudsters. During this time, they execute fraudulent

transactions, taking advantage of periods when fraud detection mechanisms may be less active or when victims are less likely to notice suspicious activities. Late evening hours often witness a decrease in monitoring and oversight from both consumers and financial institutions, making it an opportune time for fraudsters to carry out their deceptive schemes. The evening hours see an increase in online shopping activities as individuals seek convenience and leisure after work. This heightened online activity provides more opportunities for fraudsters to launch phishing attacks, engage in identity theft, or exploit vulnerabilities in payment systems, as many people are engrossed in their leisure activities and may not be as vigilant in scrutinizing transactions during these hours.

3.2.5. Feature Encoding

The conversion of categorical or textual data into numerical representation is imperative for the modelling and analysis of machine learning algorithms. Label encoding maps each category in a categorical variable to a unique integer. According to Gupta & Asha (2020), label encoding is straightforward, however, the flip side is that the numerical values can be misinterpreted by the machine learning algorithms. In their study on the impact of encoding of high cardinality categorical data to solve prediction problems. Gupta & Asha (2020) rates the best feature encoding scheme as label encoder as this has the highest performance measures amongst others.

Label encoding is suitable for big data as it requires less memory use because it represents categorical variables with integers instead of creating additional binary columns, (Hancock & Koshgoftaar, 2021). This is very useful for our base classifier, decision tree as it will allow to maintain interpretability.

	credit_card_number	merchant	category	amount	first_name	last_name	gender	street	city	state	...	trans_number	unix_time	merchant_lat	merchant_
989	4.730000e+15	605	2	67.09	229	244	0	105	55	1	...	113827	1347326909	31.318356	-88.58
333	3.540000e+15	358	2	45.34	59	420	0	742	551	14	...	511479	1353817638	38.125013	-88.89
769	4.360000e+18	447	4	108.54	314	98	1	251	619	14	...	399647	1329706091	40.768071	-91.18
667	6.760000e+11	365	2	77.35	334	283	0	364	245	16	...	618884	1337562876	40.186315	-98.33
313	4.860000e+18	330	3	30.20	27	90	0	639	703	18	...	474298	1332740118	29.970565	-90.32
748	4.210000e+18	217	4	86.67	288	23	0	677	274	14	...	734947	1342761749	38.365402	-87.64
726	4.560000e+12	176	9	5.50	73	216	1	281	300	35	...	724222	1329917418	39.609499	-84.55
838	4.650000e+15	132	6	40.82	14	244	0	423	33	47	...	549231	1340109431	46.579562	-121.77
309	2.130000e+14	576	1	73.39	150	206	1	720	638	22	...	247294	1350161485	45.842463	-83.78
414	5.620000e+11	107	2	61.46	236	45	0	327	823	7	...	360134	1332746028	38.210010	-76.64

ows × 25 columns

Figure 14: Label-encoding for categorical data

3.2.6. Feature Selection

Feature selection is a critical step in the process of building a machine learning model. It involves choosing the most relevant and informative features from the dataset to train the model. The goal is to eliminate irrelevant or redundant features, which can lead to overfitting, reduce model performance, and increase computational costs. Correlation is a widely used method for feature selection in machine learning, it involves measuring the statistical relationship between the feature variables and the target variable.

The highly correlated features will be selected for our analysis, this will help reduce noise and improve the model's performance by focusing on the most important predictors.

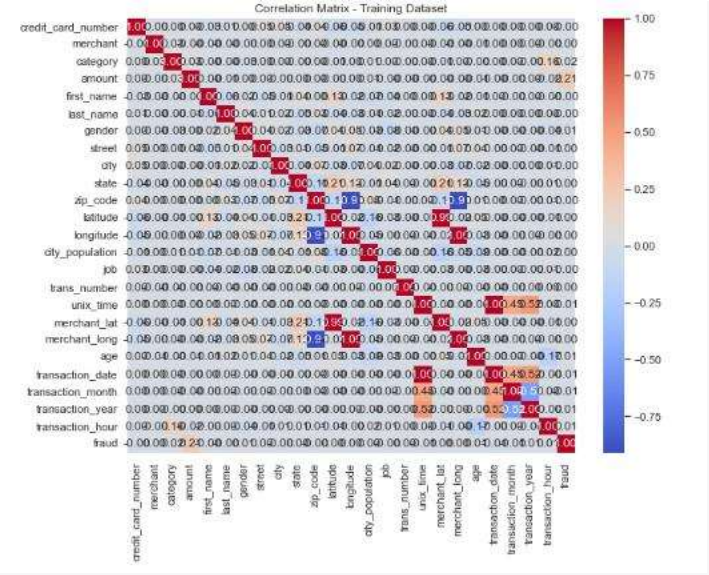

Figure 15: Correlation analysis of predictors and target variables.

The features below were selected based on the correlation with the target variable.

Strong positive correlation - Amount, category, transaction_hour, age, gender, transaction_year,

Strong negative correlation - Transaction_month, transaction_date, unix_time

Variable name	Variable Type	Description
Amount	Numerical - Continuous	The 'amount' variable represents the monetary value of each transaction and is a crucial input for training fraud detection models. It helps distinguish fraudulent transactions from legitimate ones, alongside other relevant features, using machine learning techniques.
Category	Categorical - Nominal	The 'category' variable represents the type of each transaction, such as POS or web transactions. Fraudulent activities may be more common in specific categories due to varying risk levels. Analyzing historical data allows identifying patterns in each category and flagging abnormal transactions as potentially fraudulent.
Transaction_hour	Numerical - Discrete	The variable represents the hour component of transaction timestamps. It captures the specific hour in which each transaction occurred.
Age	Numerical - Continuous	The 'age' variable represents the age of individual cardholders. It can be relevant in fraud detection to understand if certain age groups are more susceptible to fraud or if there are any age-related patterns in fraudulent activities.
Gender	Categorical - Nominal	The 'gender' variable represents the gender of individuals in the dataset. It has two categories: 'male' and 'female.' This variable be used as a feature to explore potential gender-related patterns or correlations with fraudulent transactions.
Transaction_year	Numerical - Discrete	The 'transaction_year' variable represents the year component of each transaction timestamp. It indicates the year in which the transaction occurred.
Transaction_month	Numerical - Discrete	The 'transaction_month' variable represents the month component of each transaction's timestamp. It provides information about the month in which a transaction took place. Utilizing the 'transaction_month' variable in fraud detection can help identify any potential seasonal patterns or trends in fraudulent activities across different months.
Transaction_date	Numerical - Discrete	The 'transaction_date' variable represents the day on which the transaction occurred. This variable can be useful in fraud detection to analyze temporal patterns and identify potential trends related to fraudulent activities over time.
Unix_time	Numerical - Continuous	The 'unix_time' variable represents timestamps in Unix time format, indicating the number of seconds elapsed since January 1, 1970 (Unix Epoch). It is used to record specific time points in a consistent and precise manner. In the context of fraud detection, 'unix_time' can be utilized to capture temporal patterns in transactions and their potential relationship with fraudulent activities.
Fraud	Categorical - Binary	The variable 'fraud' takes two values: 0 (representing non-fraudulent transactions) and 1 (representing fraudulent transactions). This variable serves as the target variable in the fraud detection process, where the objective is to build a predictive model that can accurately classify transactions as either fraudulent or non-fraudulent based on other features or predictors.

Table 1: Description of variables

3.2.7. Data Sampling

The dataset is highly imbalanced with less than 1% being flagged as fraudulent transaction. This could lead to overfitting of the model and could affect the accuracy of the prediction.

Figure 16: Distribution of fraudulent and non-fraudulent transactions

For an accurate prediction of our models, we used techniques such as random- undersampling, SMOTE Oversampling and SMOTE-ENN which is a hybrid sampling technique to balance the dataset.

- Random Undersampling is carried out to reduce the size of the majority class, which is the non-fraudulent transactions, to balance out the data and equalize the occurrences for each class.

```
from imblearn.under_sampling import RandomUnderSampler

# Create the RandomUnderSampler object
under_sampler = RandomUnderSampler()

# Perform undersampling on the dataset
X_train_resampled, y_train_resampled = under_sampler.fit_resample(X_train, y_train)

# Create a new DataFrame with the undersampled data
undersampled_train_data = pd.concat([X_train_resampled, y_train_resampled], axis=1)

# Print the shape of the undersampled data to verify the balance
print("Shape of the undersampled data:", undersampled_train_data.shape)
```

```
Shape of the undersampled data: (8658, 10)
```

Figure 17: Random undersampling technique

- SMOTE Oversampling, on the other hand, involves replicating instances from the minority class to balance the class distribution. This technique ensures that the classifier has sufficient data to learn from the minority class. While oversampling helps avoid information loss, it can also increase the risk of overfitting, where the model performs well on the training data but poorly on unseen data.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE


# Create the SMOTE object
smote = SMOTE(random_state=42)

# Perform oversampling on the training data
X_train_oversampled, y_train_oversampled = smote.fit_resample(X_train, y_train)

# Create a new DataFrame with the oversampled data
oversampled_train_data = pd.concat([X_train_oversampled, y_train_oversampled], axis=1)

# Print the shape of the oversampled data to verify the balance
print("Shape of the oversampled data:", oversampled_train_data.shape)
```

```
Shape of the oversampled data: (1551398, 10)
```

Figure 18: SMOTE Oversampling technique

- SMOTEENN, a combination of SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors), is a hybrid resampling technique that addresses the limitations of individual methods. SMOTE generates synthetic examples for the minority class, while ENN removes noisy examples from both classes. This approach enhances the diversity of the minority class while reducing the impact of noisy and irrelevant examples from the majority class.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from imblearn.combine import SMOTEENN  # Import SMOTEENN class for the hybrid method


# Create the SMOTE-ENN object
smote_enn = SMOTEENN(random_state=42)

# Perform SMOTE-ENN on the training data
X_train_hybridsampled, y_train_hybridsampled = smote_enn.fit_resample(X_train, y_train)

# Create a new DataFrame with the resampled data
hybridsampled_train_data = pd.concat([X_train_hybridsampled, y_train_hybridsampled], axis=1)

# Print the shape of the resampled data to verify the balance
print("Shape of the hybrid sampled data:", hybridsampled_train_data.shape)
```

```
Shape of the hybrid sampled data: (1374591, 10)
```

Figure 19: SMOTE-ENN sampling technique

3.3. MACHINE LEARNING MODELS

As stated in the previous chapter, the boosting machine learning algorithms will be implemented in this study and the performances will be compared.

3.3.1. Implementing the Adaboost Technique

- Data splitting, preprocessing, and sampling.
- Select decision tree as base classifier.
- Assign equal weights to all training samples. These weights are used to give more importance to misclassified samples in subsequent iterations.
- Train the Adaboost model by combining weak classifiers.
- Tune hyperparameters such as the number of iterations, the learning rate, and the choice of weak classifier, using the validation set to optimize model performance.
- Assess performance on testing dataset using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score).
- Visualize the AdaBoost ensemble and analyse feature importance.

3.3.2. Implementing the GBM Technique

- Data splitting, preprocessing, and sampling.
- Select Gradient Boosting Classifier library for prediction.
- Initialize the GBM model with appropriate hyperparameters, including the number of estimators (trees), the learning rate, and the maximum depth of the trees.
- Train the GBM model on the training data using the `fit` method.
- Tune hyperparameters to optimize model performance.
- Analyse feature importance scores provided by the GBM model to understand which features are most influential in making predictions.
- Evaluate the model on the testing dataset using appropriate evaluation metrics.
- create visualizations of the GBM model, such as feature importance plots.
- Implement SHAP (SHapley Additive exPlanations) to explain individual predictions.

3.3.3. Implementing the XGBOOST Technique

- Data splitting, preprocessing, and sampling.
- Install XGboost library.
- Initialize the XGBoost model with appropriate hyperparameters, such as the number of boosting rounds (trees), the learning rate, maximum depth, and more.
- Train the XGBoost model on the training data using the `fit` method.
- Use the Grid Search technique to tune hyperparameters in order to optimize model performance.
- Analyse feature importance scores provided by the XGBoost model to understand which features are most influential in making predictions.
- Evaluate the model on the testing dataset using appropriate evaluation metrics.
- Create visualizations of the XGBoost model, such as feature importance plots.

3.3.4. Implementing the LGBM Technique

- Data splitting, preprocessing, and sampling.
- Install LGBM in Python environment.
- Initialize the LGBM model with appropriate hyperparameters, including the number of boosting rounds (trees), the learning rate, maximum depth, and more.
- Train the LGBM model on the training data using the `fit` method.
- Hyperparameter tuning by adjusting learning rate, the number of trees, and the maximum depth of the trees.
- Analyse feature importance scores provided by the LGBM model to understand which features are most influential in making predictions.
- Evaluate the model on the testing dataset using appropriate evaluation metrics.

- Create visualizations of the LGBM model, such as feature importance plots or tree visualizations.

3.3.5. Implementing the CatBoost Technique

- Data splitting, preprocessing, and sampling.
- Install Catboost library.
- Initialize the CatBoost model with appropriate hyperparameters, including the number of estimators the learning rate and maximum depth.
- Train the CatBoost model on the training data using the `fit` method.
- Tune hyperparameters like learning rate, the number of trees, and the maximum depth of the trees to optimize model performance.
- Evaluate the model on the testing dataset using appropriate evaluation metrics
- Analyse feature importance scores provided by the CatBoost model to understand which features are most influential in making predictions.
- Create visualizations of the CatBoost model, such as feature importance plots or tree visualizations.

4. MODEL ANALYSIS

In the previous chapter, we implemented the boosting ensemble machine learning algorithms on resampled datasets, including random undersampled, SMOTE oversampled, and SMOTE-

ENN. We also compared their performance with that of the original dataset. Our analysis will focus on evaluating performance metrics for each of the dataset.

4.1. Model Performance on Random Undersampled Dataset

In this analysis, we assess the performance of the boosting ensemble machine learning models on a randomly undersampled dataset, which was created to address the class imbalance issue in credit card fraud detection. The classification reports provide insights into how well each model performed in terms of accuracy, precision, recall, and the F1-score.

RANDOM UNDERSAMPLED DATASET					
LEARNING MODEL			CLASSIFICATION REPORT		
		ACCURACY	PRECISION	RECALL	F1-SCORE
	ADABOOST	95%	0.09	0.92	0.16
GRADIENT BOOSTING DECISION TREE ALGORITHMS	GBM	97%	0.15	0.95	0.25
	XGBOOST	97%	0.18	0.97	0.30
	LightBOOST	97%	0.14	0.95	0.25
	CATBOOST	97%	0.15	0.97	0.26

Table 2: Model Performance on under sampled dataset.

The report above shows that Adaboost exhibits relatively lower accuracy, precision, and F1-score compared to the Gradient Boosting decision tree algorithms. However, it has a high recall, indicating that it is good at identifying positive cases (fraudulent transactions). XGBoost and CatBoost demonstrate the highest precision, which means they are better at correctly classifying positive cases while minimizing false positives for the sample data.

4.2. Model Performance on Actual Dataset

This analysis evaluates the performance of various machine learning models for credit card fraud detection using the actual dataset. The classification reports provide insights into model performance metrics.

ACTUAL DATASET					
LEARNING MODEL			CLASSIFICATION REPORT		
		ACCURACY	PRECISION	RECALL	F1-SCORE
	ADABOOST	100%	0.76	0.39	0.51
GRADIENT BOOSTING DECISION TREE ALGORITHMS	GBM	100%	0.78	0.33	0.46
	XGBOOST	100%	0.95	0.79	0.86
	LightBOOST	100%	0.92	0.74	0.82
	CATBOOST	100%	0.96	0.73	0.83

Table 3: Model Performance on Actual Dataset

The table above depicts that Adaboost shows high accuracy but comparatively lower precision, recall, and F1-score. This suggests that while it accurately classifies non-fraudulent transactions, it may miss some fraudulent ones. Gradient Boosting Machine (GBM) exhibits lower recall, indicating that it may miss a significant number of actual fraud cases.

XGBoost, LGBM, and CatBoost perform exceptionally well, with high accuracy, precision, recall, and F1-score. These models effectively detect both non-fraudulent and fraudulent transactions.

CatBoost stands out with the highest precision, making it an excellent choice when minimizing false positives is the priority.

4.3. Model Performance on Over-sampled Dataset

In this analysis, we assess the performance of boosting ensemble models for credit card fraud detection on a dataset that has been oversampled using the Synthetic Minority Over-sampling Technique (SMOTE). The classification reports provide a detailed evaluation of model performance metrics, including accuracy, precision, recall, and F1-score.

SMOTE OVERSAMPLED DATASET					
LEARNING MODEL			CLASSIFICATION REPORT		
		ACCURACY	PRECISION	RECALL	F1-SCORE
	ADABOOST	93%	0.07	0.89	0.13
GRADIENT BOOSTING DECISION TREE ALGORITHMS	GBM	96%	0.11	0.89	0.19
	XGBOOST	98%	0.17	0.91	0.29
	LightBOOST	96%	0.12	0.88	0.20
	CATBOOST	97%	0.15	0.90	0.25

Table 4: Model Performance on Over-sampled dataset

In this analysis, Adaboost exhibits relatively low precision, indicating a high number of false positives. While it has decent recall, it may still miss some fraudulent transactions. GBM improves precision compared to Adaboost but maintains a high recall. This suggests better performance in minimizing false positives while capturing more fraudulent cases. LGBM performs similarly to GBM, offering a balance between precision and recall.

XGBoost stands out with the highest precision, indicating its effectiveness in correctly classifying positive cases while minimizing false alarms. Its recall is also strong. CatBoost achieves a high recall, which is essential for identifying most fraudulent transactions, while maintaining a respectable level of precision.

4.4. Model Performance on Hybrid-sampled Dataset

This analysis assesses the performance of boosting ensemble models for credit card fraud detection using a dataset resampled with the Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors (SMOTE-ENN). The classification reports provide a comprehensive evaluation of model performance metrics, including accuracy, precision, recall, and F1-score.

SMOTE-ENN RESAMPLED DATASET					
LEARNING MODEL			CLASSIFICATION REPORT		
		ACCURACY	PRECISION	RECALL	F1-SCORE
	ADABOOST	94%	0.08	0.88	0.14
GRADIENT BOOSTING DECISION TREE ALGORITHMS	GBM	96%	0.1	0.89	0.18
	XGBOOST	97%	0.16	0.90	0.27
	LightBOOST	96%	0.11	0.87	0.20
	CATBOOST	97%	0.13	0.89	0.23

Table 5: Model performance on hybrid-sampled dataset

In the table above, Adaboost exhibits relatively low precision, indicating a high number of false positives. While it has decent recall, it may still miss some fraudulent transactions. GBM improves precision compared to Adaboost but maintains a high recall. This suggests better performance in minimizing false positives while capturing more fraudulent cases. LGBM performs similarly to GBM, offering a balance between precision and recall.

XGBoost stands out with the highest precision, indicating its effectiveness in correctly classifying positive cases while minimizing false alarms. Its recall is also strong. CatBoost achieves a high recall, which is essential for identifying most fraudulent transactions, while maintaining a respectable level of precision.

4.5. Confusion Matrix Analysis

In a Confusion Matrix, every column corresponds to predicted class instances, while each row corresponds to actual class instances. Essentially, the confusion matrix reveals how the classification model becomes "confused" when making predictions. It goes beyond just identifying classifier errors; it also provides crucial information about the nature of those errors.

In credit card fraud analysis, it is important to have the least possible false negatives and false positive cases because when the system fail to detect a fraudulent transaction, allowing it to go through (false negatives) it results in financial losses for both the card holder, merchant and issuing bank. Simultaneously, when legitimate transactions are incorrectly flagged as fraudulent (false positives), it can lead to customer inconveniences when their card is declined

for no valid reason, and repeated false positives can erode trust in the payment system and affect the reputation of the financial institution.

Upon comparing the performance metrics of the different sampled datasets mentioned earlier, it becomes evident that the actual dataset yields the most favourable results with all models having about a 100% accuracy. Consequently, we will delve deeper into the confusion matrix on the test dataset to determine the best model that is applicable for stakeholders looking at the value of the false positives and false negatives.

As seen below, the confusion matrix for the Adaboost model of the actual dataset below depicts that the model correctly identified 418 cases as fraudulent and 193,790 cases as non-fraudulent. However, the model classified 135 instances as fraud when they were not and missed 665 cases of actual fraud. This is a high number of misclassification and thus not the most reliable model for credit card fraud detection in this case.

Figure 20: Confusion matrix for Adaboost model.

The confusion matrix for the Gradient boosting model below shows that the model correctly identified 357 cases of fraud 193,823 cases of non-fraudulent transactions which is relatively higher than that of Adaboost model. On the other hand, there are 102 instances of false positive cases and 726 instances of false negative cases, shows there is room for improvement.

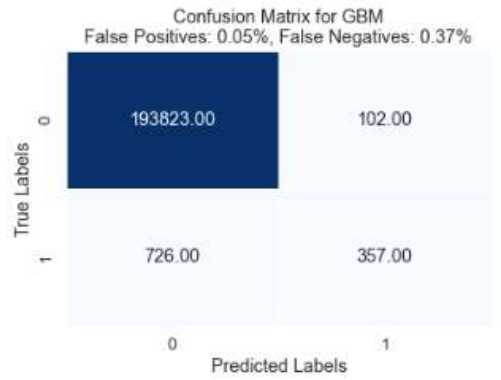

Figure 21: Confusion matrix for Gradient Boosting Model

The below result depicts the confusion matrix for Extreme Gradient Boosting (XGBoost) model. It shows that the model correctly identified 853 cases as fraudulent which is the highest compared to other models, and 193,879 cases as non-fraudulent. On the other hand, there were 41 instances of misclassified cases of fraud when they were not, and the model missed 230 cases. The XGBoost model demonstrates strong performance in identifying fraudulent transactions, with high accuracy and a reasonably balanced trade-off between precision and recall. However, the model does generate some false positives, which could lead to further investigation of those cases to reduce the number of false alarms. The model has the least false negatives at 0.12% and false positive of 0.12%, making the recommended model for stakeholders that are looking to improve operational efficiency and customer satisfaction.

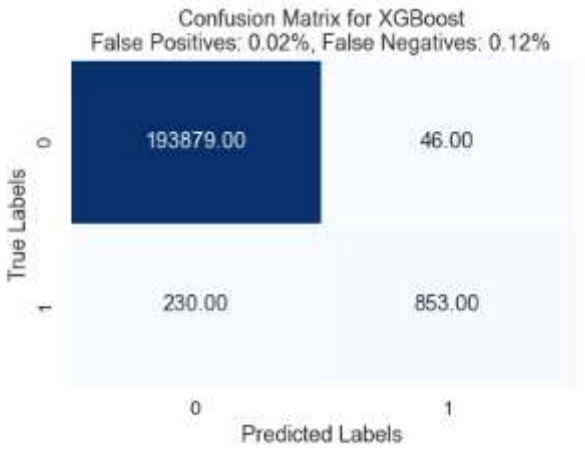

Figure 22: Confusion matrix for Extreme Gradient boosting model

The LGBM model can correctly predict 800 cases of fraud and 193,851 cases of non-fraudulent transaction. However, the model classified 74 instances as fraud when they were not and

missed 283 cases of actual fraud. This model demonstrates strong performance in identifying fraudulent transactions, with high accuracy and a reasonably balanced trade-off between precision and recall. However, like the previous analysis, it generates some false positives, indicating potential areas for further improvement to reduce the false instances.

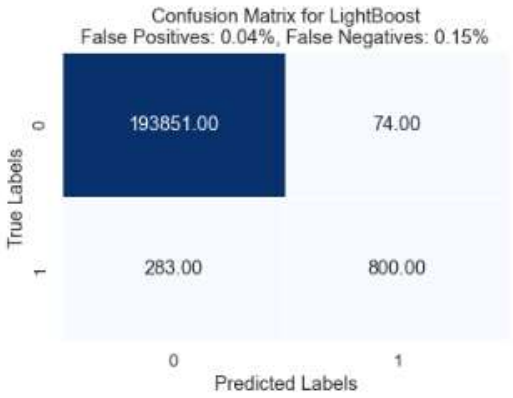

Figure 23: Confusion matrix for LGBM Model

The Catboost model has seen below, has correctly identified 795 cases of fraud and 193,890 cases of non-fraudulent transactions. The model has 35 instances of false positives and 288 instances of false negatives. The CatBoost model demonstrates strong performance in identifying fraudulent transactions, with high accuracy and a reasonably balanced trade-off between precision and recall. The Catboost model also has the least false positives at 0.02% of the analysis, making it the second-best model, if the priority for stakeholders is to improve customer experience and boost business reputation.

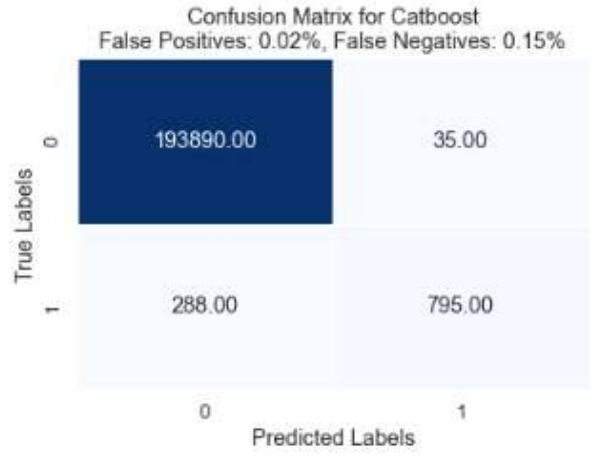

Figure 24: Confusion matrix for Catboost model

4.6. Optimal Model Interpretability

From the analysis conducted in the preceding section, we have obtained valuable insights. Specifically, we have identified XGB model as the top performing model characterized by high accuracy rate and relatively low occurrences of false positives and false negatives.

In order to gain a deeper understanding of this model and to elucidate the factors that contribute to their performance, we will explore two key aspects. Firstly, we will examine feature importance, shedding light on the relative significance of predictors. Secondly, we will implement the SHAP (Shapley Additive Explanations) framework to further interpret the model results. This endeavour aims to clarify the effects and contributions of individual predictors, facilitating comprehension for stakeholders involved in decision-making processes.

XGB MODEL

The chart below illustrates the SHAP feature importance for the XGB model, highlighting the factors that contribute the most to predictions. It reveals that the 'amount' feature holds the highest level of importance, signifying that the monetary value of a transaction has a substantial impact on fraud detection. Stakeholders should pay close attention to unusually large or small transactions, as they may be indicative of fraudulent activity. The 'category' feature ranks as the second most influential, suggesting that the types of goods or services being purchased significantly influence the determination of whether a transaction is fraudulent.

Furthermore, the 'transaction hour' feature plays a crucial role, indicating that the timing of transactions also contributes significantly to fraud detection. Stakeholders should exercise vigilance during specific hours, as these periods may be more susceptible to fraudulent activities. Age, while moderately important, should not be overlooked. Stakeholders should be aware that certain age groups may exhibit varying susceptibility to fraud, and age-related factors could impact the effectiveness of fraud detection.

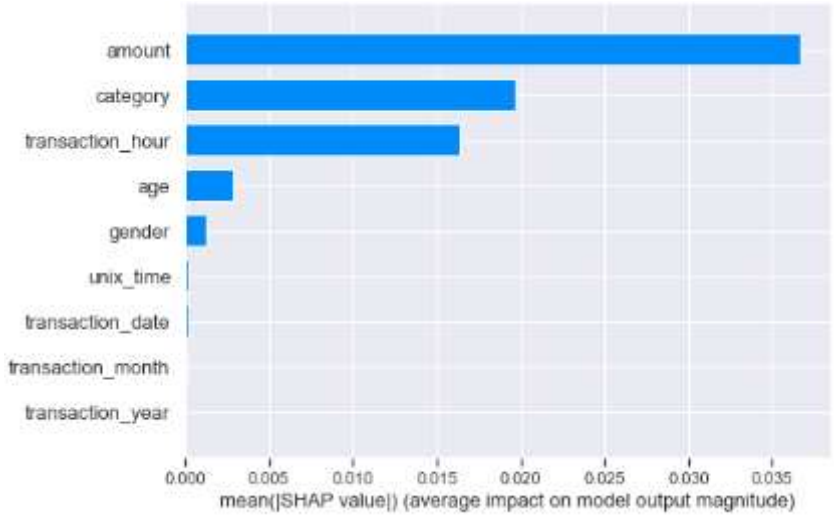

Figure 25: Feature importance of XGB Model

The SHAP interpretability plot below depicts that the 'amount' feature exhibits a notably positive impact on the predicted outcome for detecting fraudulent transactions within the dataset. This suggests that the feature tends to influence predictions in favor of the 'fraud' class (Class 1). In other words, higher transaction amounts are more likely to be associated with fraudulent transactions. Stakeholders should be attentive to transactions with exceptionally high values, as they may indicate a higher risk of fraud.

Conversely, the 'category' feature value has a prominent negative impact rate, meaning that it tends to influence predictions toward the 'non-fraudulent' class (Class 0). This implies that certain categories of goods or services are more likely to be associated with legitimate transactions. Understanding which categories have a negative impact can help in identifying transactions less likely to be fraudulent.

Meanwhile, the 'transaction hour' feature demonstrates a moderate influence that is distributed across both classes of fraud (Class 0 and Class 1). This suggests that transaction timing plays a role in predictions for both fraudulent and non-fraudulent transactions. Stakeholders should consider monitoring transaction activity during various hours as a precaution, as both classes may exhibit variability in different time periods.

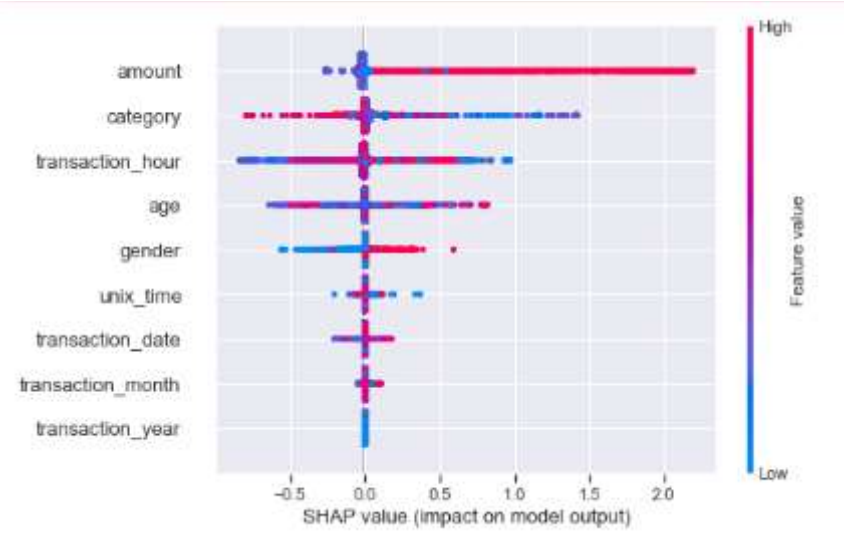

Figure 26: SHAP interpretability for XGB Model

4.7. Research Questions Addressed

The below depicts the sections where each research question and objective is addressed in this study.

Research Question	Section Addressed
How do various boosting ensemble methods impact the efficacy of fraudulent credit card transaction detection within an imbalanced dataset?	This was addressed in chapter four, the model analysis section where the result of each ensemble model is displayed based on the resampled datasets.
What are the effects of different sampling techniques on addressing class imbalance in credit card fraud detection, and how do these techniques contribute to creating a balanced dataset for model development?	This is also addressed in the chapter four, showing the influence of each sampling technique on the result of the different boosting ensemble models implemented.
What methods can be employed to enhance the interpretability of boosting ensemble models for credit card fraud detection?	This can be seen in chapter four, where the optimal model is interpreted based on feature importance and also the SHAP technique.
What strategies can be proposed to minimize classification errors and improve the practical usability of the model for fraud detection in real-world scenarios?	The objective of reducing classification errors and enhancing the model's performance in fraud detection was systematically addressed across chapters three and four. These chapters focused on implementing a range of strategies to achieve this goal, including: feature engineering, ensemble techniques, explainability amongst others.

Table 6: Research questions addressed.

5. DISCUSSION OF FINDINGS

Based on the comprehensive analysis and experimentation with various machine learning algorithms, including AdaBoost, Gradient Boosting Decision Trees (GBM), XGBoost, LightGBM (LightBoost), and CatBoost, on different datasets, including undersampled dataset, the actual dataset, SMOTE oversampled dataset, and SMOTE-ENN resampled dataset, for the purpose of credit card fraud detection, several key conclusions can be drawn:

- **Model Performance**: Our assessment of model performance, employing various classification metrics, clearly demonstrates that the Gradient Boosting Decision Tree algorithms, including XGBoost, LightGBM, and CatBoost, consistently outshine AdaBoost across all datasets in terms of accuracy, precision, recall, and F1-score. This substantiates the notion that ensemble tree-based models exhibit superior effectiveness in capturing the intricate patterns associated with credit card fraud.

  Moreover, our findings align with the research conducted by Ramani, et al. (2022), which highlights the efficiency of LightGBM, particularly when handling large datasets. Notably, our most optimal results were observed on the actual dataset, which comprises over 900,000 instances. This reinforces the value of LightGBM in real-world scenarios where dataset sizes are substantial.

  Furthermore, our study resonates with the research conducted by Chen & Han (2021) in the context of fraud detection challenges faced by the financial industry. Their investigation, conducted on a large publicly available IEEE-CIS fraud dataset, sourced from Vesta's real-world e-commerce transactions, and featuring an extensive array of features, underscores the effectiveness of the CatBoost model. In their work, CatBoost exhibited significant performance improvements in fraud detection compared to other machine learning methodologies, further underlining the potential of this algorithm in addressing complex fraud detection scenarios.
- **Dataset Resampling**: To tackle the class imbalance problem within the dataset, we applied resampling techniques, specifically SMOTE oversampling and SMOTE-ENN resampling. These techniques enhanced recall by increasing the identification of fraudulent transactions. However, they also led to an increase in false positives, resulting in reduced precision. This observation resonates with a study conducted by Bauder, et al. (2018) on Medicare fraud detection. In their research, they combined three Medicare datasets using six data sampling methods with five class ratios. Interestingly, their findings indicated that the full dataset, without any sampling, performed well across all learners. Furthermore, they observed that oversampling methods did not yield performance improvements compared to the original dataset, reinforcing the nuances of resampling techniques in different contexts.

- **Model Selection**: Among the Gradient Boosting Decision Tree algorithms, XGBoost and CatBoost demonstrated the highest F1-scores on the actual dataset, making them strong candidates for credit card fraud detection. However, LightGBM also performed impressively and is a potential choice. The overall best model, with low false positives and false negatives, is the XGBoost model on the actual dataset.

  In general, the actual dataset yields the best results because, in real-world scenarios, fraud cases are not expected to occur frequently; they are anticipated to be rare, mirroring the characteristics of the actual dataset where they constitute less than 1% of instances. This alignment with real-world expectations enhances the viability of our model's performance.

6.1. CONCLUSION AND RECOMMENDATION

The findings of the study reflect that Gradient Boosting Decision Tree algorithms provides a more reliable and accurate result for credit card fraud detection.

The CatBoost model demonstrated superior performance across multiple datasets, making it a strong candidate for implementation in real-world fraud detection systems.

In this study, the best model with relatively low false positives and false negatives is the XGB model on the actual dataset without any sampling technique.

Considering the comprehensive analysis and experimentation conducted on boosting ensemble machine learning algorithms for credit card fraud detection, several key recommendations emerge. These recommendations aim to optimize model performance and enhance the practical usability of the system.

Stakeholders and financial institution should focus on implementing a system for continuous model monitoring and retraining. As fraud patterns evolve over time, regularly update the model with fresh data to maintain its effectiveness.

It is pertinent to investigate the root causes of misclassifications and leverage this knowledge to iteratively improve the model and data preprocessing techniques.

Conducting and analysing a thorough cost-benefit analysis to assess the financial implications of model decisions, particularly false positives and false negatives will produce insights that will be used to make informed decisions about model thresholds and adjustments.

6.2. LIMITATIONS

While this research has yielded promising results and valuable insights, it is crucial to acknowledge its inherent limitations. Firstly, the study centred on a synthetic credit card fraud dataset. Although this dataset was expansive, incorporating a wide array of features and a considerable number of instances, it may not comprehensively capture the full spectrum of real-world fraud scenarios. This limitation stems from the inherent challenges of accessing actual financial transaction data due to privacy and confidentiality concerns.

Another limitation pertains to the primary mitigation strategy employed for class imbalance, which was resampling techniques. While effective to some extent, the research did not

extensively explore more advanced approaches like cost-sensitive learning or anomaly detection. These alternative methods remain ripe for further investigation and could potentially enhance the model's performance, especially in dealing with imbalanced datasets.

Furthermore, this research primarily concentrated on the development and evaluation of the model. The practical deployment of the model into real-time operational environments, taking into account factors such as latency, scalability, and continuous monitoring, presents its own set of complexities that were not fully addressed in this study. Future research should delve deeper into the operational aspects of deploying such models. Additionally, computational resources were constrained, which may have limited the extent of exploration into certain aspects, particularly hyperparameter tuning and model complexity.

In summary, despite the significant advancements made in enhancing credit card fraud detection through boosting ensemble methods, it is imperative to acknowledge these limitations. They should be viewed as opportunities for further investigation and refinement in future studies, paving the way for more comprehensive and effective fraud detection solutions.

6.3. FUTURE WORKS

Moving forward, several promising avenues for future research in credit card fraud detection emerge from the findings of this study. First and foremost is the integration of real-world transaction data into the research framework. While the synthetic dataset used in this study was comprehensive, it might not fully capture the diverse array of actual fraud scenarios encountered in the financial sector. Thus, future work should focus on securely incorporating real transaction data, balancing the need for data privacy and confidentiality with the goal of improving model accuracy.

Researchers need to consider factors such as latency, scalability, and continuous monitoring to ensure that the model functions effectively in real-world fraud detection scenarios. Future research should aim to enhance the transparency of these ensemble models, making their decision-making processes more comprehensible to end-users and stakeholders.

External factors, such as economic indicators or social engineering trends, could provide valuable context for identifying fraudulent activities. Future research may explore ways to integrate these external factors effectively into the fraud detection model.

Finally, regulatory compliance is essential for practical model deployment in the financial industry. Future research should incorporate considerations for meeting industry-specific regulatory requirements and standards. Overall, these diverse research directions aim to build upon the findings of this project and continue the pursuit of more accurate, interpretable, and practical credit card fraud detection systems.

References

AlEmad, M., 2022. *Credit Card Fraud Detection Using Machine Learning.* s.l.:s.n.

Alfaiz, N. S. & Fati, S. M., 2022. Enhanced Credit Card Fraud Detection Model Using Machine Learning. *Electronics,* 11(4), p. 662.

Alsenaani, K., 2022. *Fraud Detection in Financial Services using Machine Learning.* s.l.:s.n.

Bauder, R. A., Koshgoftaar, T. M. & Hasanin, T., 2018. Data Sampling Approaches with Severely Imbalanced Big Data for Medicare Fraud Detection. *IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI),* November.pp. 137-142.

Bhatla, T. P., Prabhu, V. & Dua, A., 2003. Understanding credit card frauds. *Cards Business Review,* 1(6), pp. 1-15.

Bhattacharya, S. & Sarkar, S., 2018. Fraud Detection in Electronic Payment Systems Using Machine Learning Techniques. *International Journal of Computer Applications,,* 180(4), pp. 32-37.

Bhattacharyya, S., Jha, S., Tharakunnel, K. & Westland, C., 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems,* 50(3), pp. 602 - 613.

Breiman, L., 1996. Bagging predictors. *Machine Learning ,* Volume 24, pp. 123-140.

Breiman, L., 2001. Random Forests. *Machine Learning volume,* Volume 45, pp. 5-32.

Cao, L., Zhang, Y. & Wang, J., 2018. Boosting-based fraud detection for online transactions.. *IEEE Transactions on Dependable and Secure Computing,* 15(5), pp. 825-838.

Chamidah, N., Santoni, M. M. & Matondang, N., 2020. The Effect of Oversampling on the Classification of Hypertension with the Naïve Bayes Algorithm, Decision Tree, and Artificial Neural Network (ANN). *Jurnal RESTI (Rekayasa Sistem dan Teknologi),* 4(4), pp. 635-641.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P., 2022. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research,* Volume 16, pp. 321-357.

Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* pp. 785-794.

Chen, Y. & Han, X., 2021. CatBoost for Fraud Detection in Financial Transactions. *IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE),* January.pp. 176-179.

Dal Pozzolo, A. et al., 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications,* 14(10), pp. 4915-4928.

Deloitte Access Economics, 2017. *Business impacts of machine learning,* s.l.: Google Cloud.

Dong, X. et al., 2020. A survey on ensemble learning. *Frontiers of Computer Science,* Volume 14, pp. 241-258.

Fernandez, A. et al., 2018. *Learning from Imbalanced Data Sets.* Cham: Springer.

FICO, no date. *Fraud Detection Machine Learning..* [Online] Available at: https://www.fico.com/en/glossary/fraud-detection-machine-learning [Accessed 23 May 2023].

Freund, Y. & Schapire, R. E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences,* 55(1), pp. 119-139.

Friedman, J. H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics,* 29(5), pp. 1189-1232.

Friedman, J., Hastie, T. & Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The annals of Statistics,* 28(2), pp. 337-407.

Galar, M. et al., 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42(2), pp. 463-484.

Gupta, H. & Asha, V., 2020. Impact of Encoding of High Cardinality Categorical Data to Solve Prediction Problems. *Journal of Computational and Theoretical Nanoscience,* 17(9-10), pp. 4197-4201.

Hancock, J. T. & Koshgoftaar, T. M., 2021. Gradient Boosted Decision Tree Algorithms for Medicare Fraud Detection. *SN Computer Science,* Volume 2, p. 268.

Han, Y. et al., 2020. Detection and Analysis of Credit Card Application Fraud Using Machine Learning Algorithms. *Journal of Physics: Conference Series,* Volume 1693, p. 012064.

Hartono, S., Salim, O., Nababan, T. & Budhiarti, E., 2018. Biased Support Vector Machine and Weighted-SMOTE in Handling Class Imbalance Problem. *International Journal of Advances in Intelligent Informatics,* 4(1), pp. 21-27.

He, H., Bai, Y., Garcia, E. A. & Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence),* pp. 1322-1328.

He, H. & Garcia, E. A., 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering,* 21(9), pp. 1263-1284.

Huang, J., Ling, C. X. & Zhang, H., 2005. Effect of Sample Size and Dimensionality on the Performance of Three Boosting Methods on Real-World Datasets. *Proceedings of the 22nd International Conference on Machine Learning (ICML),* pp. 289-296.

Ibrahim, N. B., Bin Alias, M. S. & Zin, Z. B. M., 2021. Improved sampling data Workflow using Smtmk to increase the classification accuracy of imbalanced dataset. *European Journal of molecular and clinical medicine,* 8(2).

Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B., 2020. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews,* 120(16), pp. 8066-8129.

Jha, S. & Westland, C., 2013. A Descriptive Study of Credit Card Fraud Pattern. *Global Business Review,* 14(3), pp. 373 - 384.

Ke, G. et al., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems ,* Volume 30.

Kotsiantis, S., Kanellopoulos, D. & Pintelas, P., 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering,* 30(1), pp. 25-36.

Kwak, S. K. & Kim, J. H., 2017. Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology,* 70(4), pp. 407-411.

Liang, G. & Zhang, C., 2012. An efficient and simple under-sampling technique for imbalanced time series classification. *Proceedings of the 21st ACM international conference on Information and knowledge management,* pp. 2339-2342.

Li, W., Wu, C.-S. & Ruan, S.-m., 2022. CUS-RF-Based Credit Card Fraud Detection with Imbalanced Data. *Journal of Risk Analysis and Crisis Response,* 12(3).

Li, Y., Li, H. & Yao, H., 2018. Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017. *Computational and Mathematical Methods in Medicine.*

Li, Z. et al., 2021. Local Tangent Generative Adversarial Network for Imbalanced Data Classification. *International Joint Conference on Neural Networks (IJCNN), Shenzhen, China,,* pp. 1-8.

Malik, E. F., Khaw, K. W. & Chew, X., 2022. New Hybrid Data Preprocessing Technique for Highly Imbalanced Dataset. *COMPUTING AND INFORMATICS,* 41(4), pp. 981-1001.

Mishra, A. & Ghorpade, C., 2018. Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques. *IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India,* pp. 1-5.

Mishra, S., 2017. Handling Imbalanced Data: SMOTE vs. Random Undersampling. *International Research Journal of Engineering and Technology (IRJET) ,* 4(8), pp. 317-320.

Ngai, E. W. et al., 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems,* 50(3), pp. 559-569.

Nguyen, T. H., 2014. Fraud. *The Encyclopedia of Criminology and Criminal Justice.*

Patel, R. D. & Singh, D. K., 2013. Credit Card Fraud Detection & Prevention of Fraus Using Generic Algorithm. *International Journal of Soft Computing and Engineering (IJSCE),* 2(6), pp. 2231-2307.

Penmetsa, S. D. & Mohammed, S., 2021. Ensemble Techniques for Credit Card Fraud Detection. *International Journal of Smart Business and Technology,* 9(2), pp. 33-48.

Phua, C., Smith-Miles, K., Lee, V. & Gayler, R., 2012. Resilient Identity Crime Detection. *IEEE Transactions on Knowledge and Data Engineering,* 24(3), pp. 553 - 546.

Priscilla, V. C. & Prabha, P. D., 2020. Influence of Optimizing XGBoost to handle Class Imbalance in Credit Card Fraud Detection. *Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India,* pp. 1309-1315.

Prokhorenkova, L. et al., 2018. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems,* Volume 31.

Ramani, K., Suneetha, I., Pushpalatha, N. & Harish, P., 2022. Gradient Boosting Techniques for Credit Card Fraud Detection. *JOURNAL OF ALGEBRAIC STATISTICS,* 13(3), pp. 553-558.

Rubaidi, Z. S., Ammar, B. B. & Aouicha, M. B., 2022. Fraud Detection Using Large-scale Imbalance Dataset. *International Journal on Artificial Intelligence Tools,* 31(8), p. 2250037.

Sahin, Y. & Duman, E., 2011. Detecting credit card fraud by ANN and logistic regression. *2011 International Symposium on Innovations in Intelligent Systems and Applications,* pp. 315 - 319.

Saranya, K. & Priyadharshini, R., 2021. A Comprehensive Study on Machine Learning Techniques. *International Journal of Advanced Science and Technology,* 30(3), pp. 1237-1246.

Shakya, R., 2018. *Application of machine learning techniques in credit card fraud detection.* s.l.:Doctoral dissertation, University of Nevada, Las Vegas.

Sharma, T. & Shah, M., 2021. A comprehensive review of machine learning techniques on diabetes detection. *Visual Computing for Industry, Biomedicine, and Art,* 4(30).

Snieder, E., Abogadil, K. & Khan, U. T., 2021. Resampling and ensemble techniques for improving ANN-based high-flow forecast accuracy. *Hydrology and Health System Sciences,* 25(5), pp. 2543-2566.

Strelcenia, E. & Prakoonwit , S., 2023. Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. *AI,* 4(1), pp. 172-198.

Sun, Y., Kamel, M. S. & Wang, Y., 2006. Boosting for Learning Multiple Classes with Imbalanced Class Distribution. *ixth International Conference on Data Mining,* pp. 592-602.

Taha, A. A. & Malebery, S. J., 2020. An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. *IEEE Access,* Volume 8, pp. 25579-25587.

Takacs, M., 2018. Quality Assessment of Danish government geographical data using the gradient boosting. *Journal of Geospatial Engineering,* 3(2), pp. 127-134.

Tomek, I., 1976. Two modifications of CNN.

TransUnion UK, No date. *Preventing Identity Fraud and Fraud Advice.* [Online] Available at: https://www.transunion.co.uk/consumer/credit-education/identity-theft-fraud [Accessed 25 May 2023].

UK Finance, 2022. *Over £1.2 billion stolen through fraud in 2022, with nearly 80 per cent of.* [Online]
Available at: https://www.ukfinance.org.uk/news-and-insight/press-release/over-ps12-billion-stolen-through-fraud-in-2022-nearly-80-cent-app
[Accessed 15 May 2023].

UK Finance, 2023. *The definitive overview of payment industry fraud in 2022.* [Online] Available at: https://www.ukfinance.org.uk/system/files/2023 [Accessed 25 May 2023].

Waller, M. A. & Fawcett, S. E., 2013. Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. *Journal of Business Logistics,* December, 34(4), pp. 249-252.

Willis, L. E., 2020. Deception by Design. *Havard Journal of Law and Technology,* Volume 34, p. 115.

Wilson, D. L., 1973. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics,* 2(3), pp. 408-421.

Yang, F. et al., 2022. A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis. *BMC Med Inform Decis Making,* Volume 22, p. 344.

Ye, J., Chow, J.-H., Chen, J. & Zheng, Z., 2009. Stochastic gradient boosted distributed decision trees. *Proceedings of the 18th ACM conference on Information and knowledge management,* November.pp. 2061-2064.

Zareapoor, M. & Shamsolmoali, P., 2015. Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science,* Volume 48, pp. 679-685.

Zeng, M. et al., 2016. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. *IEEE International Conference of Online Analysis and Computing Science (ICOACS), Chongqing, China,* pp. 225-228.

Zhang, X., Yaoci, H., Xu, W. & Qili, W., 2021. HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Information Sciences,* Volume 557, pp. 302-316.

Zhang, Y. et al., 2020. PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics,* 36(3), pp. 704-712.

Zhao, Y. et al., 2020. Ensemble learning predicts multiple sclerosis disease course in the SUMMIT study. *NPJ Digital Medicine,* 3(1), p. 135.

Zhu, Q., Yeh, M.-C., Cheng, K.-T. & Avidan, S., 2006. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 1491-1498.

Zhu, Y., Jia, C., Li, F. & Song, J., 2020. Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling. *Analytical Biochemistry,* Volume 593.