NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

Predicting Fraud Behaviour

A Data Mining Approach for Anti-Money Laundering

Maria Ana Mendes Correia

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREDICTING FRAUD BEHAVIOUR – A DATA MINING APPROACH FOR ANTI-MONEY LAUNDERING

By

Maria Ana Mendes Correia

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Information Analysis and Management

Supervisor/Orientador(a): Roberto André Pereira Henriques, PhD, NOVA IMS

July 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 12th of July 2024

ABSTRACT

The present dissertation evaluates the importance of using data mining techniques to prevent and detect cases of financial fraud. The most common examples of financial fraud are money laundering, credit card fraud, financial statement fraud, insurance fraud and securities and commodities fraud. A business must prevent and detect fraud behaviour in real time to avoid money losses, fines from the regulator, and exposure to financial and operational risk. Being a Bank or an Insurance, it is important to use data mining techniques to detect and prevent fraud behaviour. This study's main objective is to build a predictive model using a data mining approach and machine learning to predict money laundering in the banking sector using transaction data. The supervised learning algorithms applied to predict money laundering transactions are Logistic Regression, Neural Networks, Decision Trees, Random Forests, Light Gradient Boost and Ensemble. The dataset used in this study was highly imbalanced, and it is necessary to apply an oversampling technique that combines K-means clustering with SMOTE. The empirical results show that the Light Gradient Boost is the model with the best performance, showing a strong discriminatory power (AUC=99,9% and Gini=0,998), a strong precision (98,4%) and recall (96,4%). It achieved the highest value of f1-score (97,4%), showing that the model correctly identifies a high number of fraudulent transactions while minimizing the false positives and false negatives. This study proves that by monitoring and analyzing transaction data, fraudulent transactions can be predicted with high levels of success achieved. It also presents a strong evidence that data mining techniques can continuously be used to detect cases of fraud behaviour, especially cases of financial fraud in the Banking sector.

KEYWORDS

Fraud Behaviour; Financial Fraud; Money Laundering; Supervised Learning; Imbalanced data; Oversampling; Data Mining

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AML	Anti-Money Laundering
AUC	Area Under the Curve
DM	Data Mining
DT	Decision Tree
EDA	Exploratory Data Analysis
FT	Financial Terrorism
FF	Financial Fraud
KNN	K-nearest neighbors
LGB	Light Gradient Boost
LR	Logistic Regression
ML	Money Laundering
MSE	Mean Square Error
NN	Neural Network
PEP	Political or Exposed Person
RCA	Relative or Close Associate
RF	Random Forest
ROC	Receiver Operating Characteristic
ROS	Random Over Sampling
RUS	Random Under Sampling
SMOTE	Synthetic Minority Oversampling Technique

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Fraud has always existed in our society and has been increasing over the years, arming many businesses' health and solid structure (Albashrawi & Lowell, 2016). Since the 1980s, a set of unofficially or even nonregistered economic activities has been known, and it has also been recognized that this phenomenon is global and is not dependent on the current economic model (Pimenta, 2009). This leads to an important question: what originates fraud? The source of fraud is human behaviour, and many motives contribute to an individual's willingness to commit illegal activities, such as greed, financial need and pressure from their families, pathological desire and tendency to beat the system (Egiyi et al.,2023).

When fraud behaviour happens in the financial world, we are dealing with financial fraud. Financial fraud can be defined as an intentional act of deception involving financial transactions for personal gain. The most common examples of financial fraud are money laundering, credit card fraud, financial statement fraud, insurance fraud and securities and commodities fraud (West & Bhattacharya, 2016).

Financial fraud occurs mainly in the banking and insurance sector since these types of businesses deal with huge amounts of money transactions, which makes it difficult to control the origin of the money. Money laundering is one type of financial fraud that consists of taking cash earned from illegal activities and making the cash appear to be gained from a legal activity (Hilal et al., 2022). This study will focus on this type of financial fraud.

The massive adoption and constant day-to-day usage of online shopping, digital payments, home banking, mobile apps and transactions through payment cards are contributing to the increase in fraud behaviour and, therefore, money laundering activities. With the evolution of modern technology, fraudsters are also evolving in the same length and adapting their tactics to exploit vulnerabilities in existing anti-money laundering systems (Pinzón et al., 2023).

In light of this, a business needs to prevent and detect fraud behaviour in real time to avoid money losses and fines from the regulator and to avoid exposure to financial and operational risk. Given the high volume of banking and transactional data, banks have been using data mining techniques to detect money laundering patterns and prevent financial fraud (Salehi, Ghazanfar & Fathian, 2017).

The literature on this topic mainly concentrates on creating predictive models for financial fraud in the Banking and Insurance sector. Currently, the available studies and the application of data mining techniques tend to be higher for credit card fraud when compared to money laundering detection. The supervised algorithms most used are the following: Neural Networks, Logistic Regression, Support Vector Machine, Naive Bayesian, Decision Trees and Random Forest. However, there is no comprehensive view of the best data mining technique for fraud detection (Albashrawi & Lowell, 2016).

1.2. STUDY RELEVANCE AND IMPORTANCE

Financial fraud has been a massive concern for organizations because of the considerable impact and devastation illicit activities bring to the business. Billions of dollars are lost during the year. In

Albashrawi & Lowell (2016), we have an example of an American Bank that paid over $16.5 billion to solve a financial fraud case.

According to Mills (2017), the total cost of fraud, money laundering or terrorism financing and the set of risks that a financial institution faces after a fraud attack or other illegal activity go far beyond the losses themselves. An average organization may lose 5% of annual revenues to fraud, while the global loss is estimated at approximately 3,7 trillion dollars.

A more recent economic study by PwC in 2020 reported that 49% of the surveyed companies had been victims of financial fraud. About 26% of the companies lost between 50.000 and 1 million dollars through fraud over the last 24 months, and 3% reported losses of over 5 million.

According to the United Nations Office on Drugs and Crime, the global value of laundered money in one year ranges between 500 billion dollars to 1 trillion dollars (Le Khac & Kechadi, 2010).

Considering these results, an organization needs to use predictive analytics to detect and predict fraud behaviour. This is a severe matter for financial institutions, mainly because Banks and Insurance companies are exposed to several risks when involved in illegal activities: reputation, operational, concentration, and legal. Employing data mining techniques will aid financial institutions in detecting, predicting and preventing fraud cases, such as money laundering (Lo & Li, 2016).

In continuation of the abovementioned, when a company invests in fraud detection, it will benefit from the following advantages: prompt reaction to fraudulent activities, exposure reduction to fraudulent activities, reduced economic damage caused by fraud, identification of the vulnerable accounts more exposed/susceptible to fraud and increased trust and confidence of the organization's shareholders.

1.3. Study Objectives

The main goal of this study is to create a predictive model in the Financial Industry, specifically in the Banking sector, to detect suspicious money laundering cases. The model should use historical transaction data to predict if a transaction is fraudulent or non-fraudulent.

The main objectives of this study are in line with the following statements:

- Identification of the most relevant variables to build the predictive models;
- Understand if outliers should be removed when dealing with fraudulent transactions. By excluding outliers, we can exclude valuable information for the prediction;
- Understand which oversampling technique should be applied when dealing with an imbalanced dataset;
- Understand which are the best algorithms to apply when it comes to predicting fraud behaviour;
- Identification of the most used algorithms to predict money laundering activity; and,
- Identification of the model with the best performance.

Over the past years, several studies have shown data mining techniques and applying different algorithms to predict financial fraud: Neural Networks, Logistic Regression, Support Vector Machine, Naive Bayesian, Decision Trees, Random Forest and Ensemble. Each algorithm was able to show its advantages, and their performance varied depending on the type of fraud being predicted (e.g., Credit Card fraud, Financial Statement fraud, Insurance fraud or Money Laundering), on the approach

presented by the author and if the data source considered in the study had problems of imbalanced data. Recent studies show the usage of Boosting algorithms to predict fraudulent transactions, achieving high-accuracy results (Gamini et al., 2021).

In this study, we highlight the importance of using data mining techniques to prevent and detect cases of money laundering in the banking sector, proposing the creation of six different supervised learning algorithms to predict fraudulent transactions: Logistic Regression, Neural Networks, Decision Trees, Random Forest, Light Gradient Boost and Ensemble.

In this dissertation, it's important to mention that the expression "fraudulent transaction" means that the transaction originated in illegal activity such as money laundering, and "non-fraudulent transaction" means that it is a legal transaction.

The remainder of this dissertation is organized as follows. In the second chapter, is presented the concept of Fraud and Money Laundering. The third chapter summarizes related work regarding Credit Card fraud, Financial Statements fraud, Insurance Fraud and Money Laundering detection and prediction. In the fourth chapter, the methodology of this dissertation is presented. The fifth chapter shows the results of each model and a comparison between each model's performance. The six chapter presents the conclusions of this study, followed by chapter seven presenting the identified limitations and recommendations for future work.

2. THEORETICAL BACKGROUND

2.1. THE CONCEPT OF FRAUD

Fraud is intentional deception to secure unfair or unlawful gain or to deprive a victim of a legal right. It is a criminal act of misleading others to damage them or their business to get something of worth for their own advantage (Mukherjee, Mukherjee, & Nath, 2016).

Financial fraud can be broadly defined as an intentional deception involving financial transactions for personal gain. Usually, when referring to financial fraud, we include the following examples: money laundering, credit card fraud, financial statement fraud, insurance fraud and other fraud techniques (Hilal et al., 2022).

This concept is not new; over the past years, there have been countless examples of fraud behaviour in the business environment. The development of technologies, the Internet of Things, the increasing usage of mobile apps, e-commerce, and the creation of BitCoins are some examples that contribute to new ways of doing fraudulent activities.

According to an economic crime survey by PwC in 2018, fraud is a billion-dollar business. It is increasing yearly: almost half (49%) of the 7,200 companies surveyed have experienced fraud. Most fraud activities involved cell phones, tax return claims, insurance claims, credit cards, supply chains, retail networks, and purchase dependencies.

Fraudulent activity is a high-cost threat that can compromise the integrity of an organization. We also need to remember that this behaviour is not exclusively external to the organization. Fraud can take the form of internal activity (internal fraud), such as an employee modifying financial records, or can arise from an external threat (external fraud), such as customer credit card fraud. External fraud can be divided into three main profiles: the average offender, the criminal offender and the organized crime offender (Phua et al., 2010). The last two profiles tend to have higher fraud activity.

2.2. THE CONCEPT OF MONEY LAUNDERING

Money Laundering is the process of making illegal income appear legal. This is the process by which criminals attempt to hide the true origin of their criminal activity. Through money laundering, criminals try to convert monetary proceeds from illegal activities into clean funds using a legal medium such as significant investment funds hosted in investment banks (West et al., 2016) (Hilal et al., 2022).

Money Laundering can be classified as a dynamic three-stage process:

- Placement: the movement of cash from its source. The source is usually disguised or misrepresented to avoid raising suspicion. This is followed by placing it into circulation through financial institutions, casinos, shops and other types of business;
- Layering: the nature of this step is to make it more difficult to detect and uncover laundering activity for the law enforcement agencies;
- Integration: the movement of previously laundered money into the economy mainly through the banking system, appearing to be normal business earnings.

Money laundering is a serious threat to financial institutions and each country. Every country should worry about Money Laundering because this type of fraud influences political and economic stability. Therefore, financial regulators require that financial institutions implement techniques and procedures to prevent and detect money laundering and other illegal activities such as financing of terrorism, for example.

3. LITERATURE REVIEW

Over the past years, there has been an increase in articles about the prediction of fraud behavior by applying data mining or machine learning methods. Most of the articles cover predictions related to money laundering, credit card fraud, financial statements fraud and insurance fraud.

Different approaches and algorithms were used, tested, and evaluated depending on the type of fraud to be detected. Some algorithms tend to be more accurate, performing better in one kind of fraud when compared to another type (West & Bhattacharya, 2016).

Below are the main articles on data mining techniques applied to fraud detection and prediction.

3.1. CREDIT CARD FRAUD

Brause et al. (1999) started applying data mining techniques to detect credit card fraud. Their study shows how advanced data mining techniques and Neural Network algorithms can be combined successfully to obtain a high fraud coverage and a low false positive rate. For their analysis, they used a sample set of 5.850 fraud transactions and 30.000 legal transactions (Brause, Langsdorf & Hepp, 1999).

Yeh & Lien, to predict credit card fraud, used the following algorithms: Logistic Regression, K-Nearest Neighbor, Discriminant Analysis (Fisher's Rule), Neural Networks, Naive Bayesian and Decision Trees for this study payment data (cash and credit card issue) were analyzed from a vital bank in Taiwan. The targets were bank credit card holders. The dataset had 25.000 observations, with 5.529 transactions labelled as fraud – cardholders with default payment. Among the six algorithms, the artificial Neural Network was the best model in classification accuracy with 54%, followed by Decision Trees with 53,6%. In this study, they also wanted to estimate the actual probability of default by applying the Sorting Smoothing Method (SSM). In this method, each predictive model was compared with the predicted probability. Neural Network was also the model with the best predictive accuracy of probability of default, showing values of R^2 equal to 0,9647 (that is close to 1), regression intercept equal to 0,0145 (close to 0) and regression coefficient equal to 0,9971 (close to 1). This study shows that Neural Networks should be employed to score clients instead of other data mining techniques (Yeh & Lien, 2009).

In 2011, Bhattacharyya et. al (2011), evaluated three data mining approaches: Support Vector Machines (SVM), Random Forests and Logistic Regression. This study aimed to examine the performance of Random Forests and SVM, together with the Logistic Regression for credit card fraud identification. The authors used real-life data on credit card transactions to train the models. This data was obtained from international credit card operations and has 13 months (from January 2006 to January 2007) of about 50 million (precisely 49.858.600 transactions) credit card transactions. The authors also had the problem of the dataset being highly imbalanced, which means that the proportion of fraudulent transactions is lower (<1%) compared to the proportion of non-fraudulent transactions. The authors stated that random under-sampling of the majority class is generally better at solving this problem, so using this option, they obtained training datasets of different proportions of fraud cases: 15%, 10%, 5% and 2% of fraudulent transactions. A data set with only 0,5% of fraud cases and a much lower fraud rate was also tested to evaluate the performance of the models when dealing with a reality when the proportion of fraudulent transactions is low. The following performance metrics were

analysed to choose the champion model: overall accuracy, weighted accuracy, sensitivity, specificity, precision, f-measure, AUC performance measure and geometric mean. Logistic Regression showed particularly low accuracy when the fraud rate in the training data was at the lowest level (2% fraud). Random Forests showed the best performance, followed by Logistic Regression and then SVM; for the training dataset with the lowest fraud rate, however, SVM surpasses Logistic Regression and performs similarly to Random Forests. This pattern, with SVM matching the performance of Random Forests when trained with the lowest proportion of fraud cases in the data, is also seen for the other measures. Regarding precision, Random Forests showed the highest performance; SVM and Logistic Regression are similar, except when dealing with 2% fraud, where SVM's performance approaches that of Random Forests. On the F-measure, which reflects a trade-off between accuracy on fraud cases and precision in predicting fraud, Random Forests showed better performance; with the lowest fraud rate in the training data (with 2% fraud), SVM performs similarly to Random Forests, with Logistic Regression far lower. Performance on both the G-mean and weighted accuracy measures, which consider accuracy on fraud and non-fraud cases, is similar to that for sensitivity. In conclusion, random forests performed better than the other techniques in all performance measures (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011).

In the same year, Dharwa & Patel proposed a hybrid approach for fraud detection of online financial transactions, containing data mining techniques, artificial intelligence and statistics in a single platform. The proposed model, the Transaction Risk Score Generation Model (TRSGM), consists of five components: Density-Based Scan Algorithm (DBSCAN), Linear Equation, Rules, Data Warehouse and Bayes Theorem. The data used in this work was from an online shopping firm. The authors concluded that the model is flexible; new rules can be added, and weightage is dynamic. In addition, the fact that Bayes Theorem is used gives the advantage of adapting to the changing behaviour of the customer. The fact that the model also uses unsupervised learning allows the detection of new types of fraud (Dharwa & Patel, 2011).

In 2018, Hordri et al. studied how to handle class imbalance in credit card fraud using resampling methods. The authors used the most common resampling methods: Random Over-Sampling (ROS), Random Under-Sampling (RUS) and Synthetic Minority Over-Sampling Technique (SMOTE). They used a publicly available dataset of 284.807 total transactions made in September 2013 by European cardholders. The dataset was highly imbalanced, having 492 fraud transactions. For each resampling method, four algorithms were applied: Naive Bayes, Linear Regression, Random Forests and Multilayer Perceptron. The performance of the classifiers was compared in terms of the following metrics: Sensitivity, Specificity, Accuracy, F-Measure and Area Under Curve (AUC). Random Forests was the model that performed well in the three resampling methods, having a higher accuracy when compared to the other classifiers. When comparing only the resampling methods, the SMOTE was quite effective. ROS also presented convincing results when compared to SMOTE. The authors stated that it would be interesting to compare the classification techniques used in this study with others like Support Vector Machine (SVM), Neural Networks (NN) and Genetic Algorithm (Hordri, Sophiayati, Firdaus & Mariyam, 2018).

In the same year, Rajora et al. proposed a comparative performance of ten machine learning algorithms to predict credit fraud. Something interesting is that the variable Time was introduced in the dataset. The algorithms' performances were evaluated with and without the time variable. The

authors also encountered the problem of dealing with an imbalanced dataset, having more normal transactions than fraud transactions (248.807 credit card transactions and only 492 labelled as fraud transactions). Therefore, to solve the imbalance problem, the authors applied the under-sampling approach, removing data from the majority class until their amount was roughly equal to the minority class. The classification algorithms analysed were the following: Naive Bayes, Decision Trees, Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Ensemble. In the Ensemble algorithm, the authors analysed the following methods: Bagging, Random Forest, Adaptative Boost, Gradient Boosting Machines and Gradient Boosted Regression Trees. By examining the results, Decision Trees, Naive Bays and SVM performed better with the Time variable. However, the performances without Time were slightly better for the Adaptative Boost and Bagging. All the other performances of the remaining algorithms had no difference between the two options. Considering all attributes, Logistic Regression and SVM had better performance than the others achieving 0,94 AUC score for the overall performance comparison. The five approaches perform better than the five individual models regarding the Ensemble algorithms. Random Forest, Adaptative Boost and Gradient Boosted Regression Trees were the ones to stand out with 0,94 AUC scores (Rajora, Li, Jha, Bharill, Patel, Joshi, Puthal & Prasad, 2018).

In 2021, Gamini et al. used only Boosting algorithms to predict fraudulent transactions with credit cards. Extreme Gradient Boost (XGBoost), CatBoost and Stochastic Gradient Boosting were the three chosen algorithms. These algorithms were implemented on a dataset composed of 248.007 transactions, each classified as fraud or not a fraudulent transaction. The authors used the Confusion Matrix, Precision and Recall metrics to evaluate the model's performance. All algorithms achieved a high accuracy (> 90%); nonetheless, the Catboost algorithm was the best model, with an accuracy equal to 92% and a recall equal to 1. In this study, the dataset was also imbalanced. Although the sampling method to address the imbalanced problem is not specified, since the authors only indicated the application of data mining techniques to deal with the imbalance, they were able to achieve a successful performance of the models (Gamini, Prathima & Yerramsetti, Sai & Darapu, Gayathri & Pentakoti, Vamsi & Prudhvi, Vegesna, 2021).

The following year, Bhuiyan et al. presented an approach to handling class imbalance when predicting credit card fraud. The dataset used in this study was a European cardholder’s dataset collected from the ULB group, where only 1,7% of the transactions were labelled as fraud cases. To overcome the class imbalance in the dataset, the authors applied different sampling techniques such as Under-Sampling, Over-Sampling, SMOTE and AdaSyn (Adaptive Synthetic). The resampled dataset was then used for classification using machine learning algorithms like Decision Trees, Random Forest, K-Nearest Neighbours, Logistic Regression and Naive Bayes. They used Recall, F1-score, accuracy, precision and error rate to evaluate the model's performance. Analysing by type of sampling technique, the Logistic Regression achieved the highest accuracy result with 99,94% after under sampling technique, and the Random Forest achieved the highest accuracy result with 99,96% after over sampling technique. Among the four sampling techniques, SMOTE demonstrated a strong performance (Bhuiyan, Khatun, Taslim & Hossain, 2022).

3.2. Financial Statements Fraud

In 2007, Kotsiantis et al. explored the effectiveness of machine learning to detect firms that issued fraudulent financial statements (FFS). They used the following algorithms to predict fraudulent financial statements: Decision Trees, Neural Networks, Bayesian Networks, Support Vector Machine, Nearest Neighbor and Ensemble classifier. The algorithms were trained using a sample of 164 Greek firms on the Athens Stock Exchange, where 41 were identified as fraud and 123 as non-fraud. The event rate for fraud is equal to 25%. By analysing the results, the best model in terms of accuracy was the Ensemble classifier. Using a stacking variant methodology, the algorithm presented an accuracy of 95,1% in the validation set. Decision Trees also achieved a good performance, having an accuracy of 91,2% in the validation set (Kotsiantis, Koumanakos, Tzelepis & Tampakas, 2007).

In the following year, Liu studied the differences and similarities between the two models: prediction of business failure and prediction of fraudulent financial reporting. The algorithms applied were Neural Networks, Logistic Regression and Decision Trees. The authors used a sample with data obtained from the Taiwan Economic Journal data bank and Taiwan Stock Exchange Corporation about some Taiwanese firms that experienced financial problems or were accused of fraudulent reporting in 2005. The Logistic Regression was the one with the best performance in both models, having an accuracy of 99,05% in predicting business failure and 96,5% accuracy in predicting fraudulent financial reporting (Liu, 2008).

Deng and Mei used an unsupervised learning approach to predict fraudulent financial statements in 2009. This study shows that unsupervised algorithms can also be used to detect fraud, achieving good results. The dataset comprised 100 Chinese firms, and 47 financial ratios were chosen as variables. They designed a clustering model called V-KSOM, combining Self-Organization Map (SOM) and K-means clustering. Analyzing the experimental results, the best accuracy rate was equal to 89%, also having a Silhouette index (a metric that measures the consistency of the clusters, varying between -1 and 1) equal to 0,2707 (Deng & Mei, 2009).

3.3. Insurance Fraud

In 2005, Viaene et al. studied the detection of automobile fraud using Neural Networks. The empirical evaluation is based on a dataset of 1399 closed automobile insurance claims from accidents in Massachusetts, USA, during 1993, and for which information was collected by the Automobile Insurance Bureau (AIB). The Neural network results were compared to the Logistic Regression and Decision tree classifiers (Viaene, Dedene & Derrig, 2005).

Yang & Hwang proposed a data mining framework that detects fraudulent and abusive behaviour in healthcare insurance. They used a real-world dataset of 1812 medical claims submitted to Taiwan's Bureau of National Health Insurance (BNHI). Each medical claim was labelled as a fraud or regular case. Their first approach was to use the structure pattern algorithm to find patterns in the data. Each structure pattern was considered a feature. They used probabilistic reasoning to reduce feature space and to minimize the amount of lost information. Their second approach was to build the detection model using the features of the structure pattern discovery algorithm. The authors adopted the Classification Based on Associations algorithm (CBA). To evaluate the model performance, sensitivity and specificity were the selected measures. The authors outlined that their detection model proved to

be more efficient and capable of identifying fraudulent and abusive cases that are not detected by manually constructed detection models (Yang & Hwang, 2006).

In 2011, Thiprungsri & Vasarhelyi used an unsupervised learning algorithm to detect anomalies. Cluster analysis was applied in the accounting domain, specifically in detecting anomalies in the audit field. The data used in this study was from a group life claims business of a major US insurance company. The dataset contains 40.080 records of group life claim payments issued in the first quarter of 2009. Using the K-means clustering procedure, eight clusters were formed. By analysing and comparing all clusters, the authors identified 169 outliers and 568 claims as possible anomalies because these claims presented lower than 0,6 probabilities of belonging to the cluster they were assigned. The authors concluded that cluster analysis is a helpful audit technology and a good option for fraud and anomaly detection (Thiprungsri & Vasarhelyi, 2011).

In 2014, Rodrigues & Omar identified the most economical model to detect suspected fraud cases in automobile insurance. Their objective was to find a model to decrease insurance companies' costs with fraud. The dataset had alleged fraud cases between 1994 and 1996, with 15.421 instances in total. Each case was classified as fraud or legal. The dataset was divided into two partitions to train and test the following classifiers: Decision Trees, Naive Bayes, Neural Networks, SVM and Ensemble. The training partition has automobile claims between 1994 and 1995, and the testing partition has automobile claims from 1996.

Nevertheless, the dataset was imbalanced, meaning the classes were not distributed in the same quantity. The authors solved this problem by using the subsample generation. The confusion matrix was extracted using the testing dataset to calculate the cost model value for each tested classifier. The champion model combined all models – the ensemble method, using an average vote function decision (Rodrigues & Omar, 2014).

Hassan & Abraham proposed an insurance fraud detection method for imbalanced data distribution. They used Decision Trees, SVM and Neural Networks on data partitions derived from under-sampling, with replacement and without replacement, of the majority class and merging it with the minority class. The authors used the ten-fold cross-validation method to choose the best under-sample partition. The results demonstrated that Decision Trees were the algorithm that performed better (Hassan & Abraham, 2016).

3.4. MONEY LAUNDERING

In 2005, Tang & Yin developed an intelligent data-discriminating system for anti-money laundering based on the Support Vector Machine (SVM) algorithm. The authors first used statistical learning theory (SLT) to improve the embarrassment of anti-money laundering, and then they used the SVM algorithm to detect unusual behaviour. Also, in 2011, Keyan & Tingting presented an improved support-vector network model for anti-money laundering. Their objective was to find the optimal SVM classifier parameters using the cross-validation method based on the highest classification accuracy rate.

Lv, Ji & Zhang proposed the Radial Basis Function (RBF) Neural Network model to track ML activity. The training set comprises one million records transactions over eight months from 6000 accounts. The available attributes include the client number, client name, capital account, certificate number of the

client, transaction date, business types affiliated with the transaction account, code of the transaction area, sum of transaction amount, transaction time, transaction currency, transaction types and frequencies of transaction. The model presented in this paper could reach high correction rates, reducing false positive rates and enhancing positive rates. The authors concluded that the RBF Neural Network model is feasible and effective when detecting money laundering (Lv, Ji & Zhang, 2008).

In 2010, Le Khac & Kechadi applied a knowledge-based solution that combines data mining and computing techniques to detect money laundering patterns. They tested Clustering (K-means), Neural Networks, Genetics Algorithm and Heuristic Algorithm using transactional datasets from six funds of CE Bank, an international investment bank. The dataset had approximately 10 million transaction records. The authors concluded that their approach could satisfy the needs of detecting money laundering. They also stated that their solution can improve the performance of the current CE bank's solution in terms of running time (Le Khac & Kechadi, 2010).

Liu et al. used Decision Trees to identify money laundering activities. The authors used the first cluster analysis by combining the BIRCH and K-means algorithms to research and identify typical money laundering patterns and rules. After that, the authors applied the Decision Tree algorithm to detect abnormal transaction data (Liu, Qian, Mao & Zhu, 2011).

In 2016, Khalaf & Khamesy aimed to answer the question "who are money laundering and who are not?". They used a data mining framework based on testing and evaluating four types of Neural Networks: Multi-Layer Perceptron Neural Networks (MLP), Probabilistic Neural Networks (PNN), Radial Basis Function (RBF) and Linear Neural Networks (LNN). To train the models, they used data from regular bank records. The Linear Neural Network was the model with the best performance, with 80% of data correctly classified and an error classification rate equal to 36%. Based on the results obtained, they concluded that the model could be used in other financial transactions (Khalaf Ahmed Allam El-Din & El Khamesy, 2016).

Sain and Puri presented different methodologies to detect suspicious accounts involved in money laundering activities. Their objective was to review research conducted in the field of fraud detection with an emphasis on data mining techniques to detect money laundering. By analysing their research, one can conclude that clustering methods and prediction using Neural Networks, SVM, Decision Trees, and Bayesian networks are the most used. However, some studies have implemented social network analysis, time series, Euclidean distance, sequence miner algorithm, money transfer analysis, and behavioural patterns (Sain & Puri, 2018).

In a more recent study, Lokanan aimed to build five machine-learning algorithms to detect the probability of money laundering in banks. The algorithms were Naive Bayes, Logistic Regression, Random Forest, Catboost algorithm and artificial Neural Networks. The data used for the study came from a simulation of money laundering activities in the Middle East banks based on a real dataset. The dataset contains information about the type of transaction, level of crime, currency amount, transaction date, transaction time, kind of money laundering, and the target variable that indicates whether the transaction is classified as money laundering. The authors used the accuracy, precision, recall, and F1-score measures to evaluate the model's performance. Out of the five models, the Naive Bayes and Random Forest were the best-performing models with an equal value of 77,46% accuracy (Lokanan, 2022).

3.5. Summary of the Literature Review

The table below provides a quick and concise summary of all the reviewed articles.

Table 1 - Review of Prior Research about Fraud Prediction

Reference	Title	Fraud Type	Methods	Findings
(Brause, Langsdorf & Hepp, 1999)	Neural Data Mining for Credit Card Fraud Detection	Credit Card	Neural Network	Advanced data mining techniques and Neural Network algorithm can be combined successfully to obtain a high fraud coverage and a low false positive rate.
(Yeh & Lien, 2009)	The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients	Credit Card	Logistic Regression K-Nearest Neighbor Discriminant Analysis Neural Networks Naive Bayesian Decision Trees	The artificial Neural Network was the best model in terms of classification accuracy (54%), followed right after by Decision Trees (53,6%). This study shows that Neural Networks should be employed to score clients instead of other data mining techniques.
(Bhattacharyya, Jha, Tharakunnel, & Westland, 2011)	Data mining for credit card fraud: A comparative study	Credit Card	SVM Random Forests Logistic Regression	Highly imbalanced dataset. Random under-sampling of the majority class was used to obtain training datasets of different proportions of fraud cases: 15%, 10%, 5% and 2% of fraudulent transactions. In terms of precision, Random Forest showed the highest performance. SVM and Logistic Regression were similar, except when dealing with 2% fraud, where SVM's performance approaches that of Random Forest.
(Dharwa & Patel, 2011)	A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction	Credit Card	Density-Based Scan Algorithm (DBSCAN) Linear Equation Bayes Theorem	Proposed model: Transaction Risk Score Generation Model (TRSGM) consisting in 5 components -DBSCAN, Linear Equation, Rules, Data Warehouse and Bayes Theorem. The model is flexible: using Bayes Theorem gives the advantage of adapting to the changing behaviour of the costumer and using an unsupervised learning gives the opportunity to detect new types of fraud.
(Hordri, Sophiayati, Firdaus & Mariyam, 2018)	Handling Class Imbalance in Credit Card Fraud using Resampling Methods	Credit Card	Naive Bayes Linear Regression Random Forests Multilayer Perceptron	How one can handle class imbalance in credit card fraud using resampling methods: ROS, RUS, SMOTE. Random Forests was the model that showed a good performance in the three resampling methods, having the higher accuracy. Comparing only the resampling methods, the SMOTE was quite effective. ROS also presented convincing results when compared to SMOTE.
(Rajora, Li, Jha, Bharill, Patel, Joshi, Puthal & Prasad, 2018)	A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance	Credit Card	Naive Bayes Decision trees Logistic Regression K-nearest neighbour SVM Ensemble	Working with an imbalanced dataset, using under-sampling to solve the imbalanced problem. In the Ensemble algorithm, the authors analysed the following methods: Bagging, Random Forest, Adaptative Boost, Gradient Boosting Machines and Gradient Boosted Regression Trees. These five approaches had better performance than the five individually models.
(Gamini, Prathima & Yerramsetti, Sai & Darapu, Gayathri & Pentakoti, Vamsi & Prudhvi, Vegesna, 2021)	Detection of Credit Card Fraudulent Transactions using Boosting Algorithms	Credit Card	XGBoost Algorithm Catboost Algorithm Stochastic Gradient Boosting (SGB) Algorithm	This paper also shows how machine learning algorithms can continuously assist in finding fraud detection in credit card transactions, in this case using only Boosting algorithms. The Catboost algorithm was the one with the highest prediction accuracy (92%).
(Bhuiyan, Khatun, Taslim, & Hossain, 2022)	Handling Class Imbalance in Credit Card Fraud Using Various Sampling Techniques	Credit Card	Decision Tree Random Forest K-Nearest Neighbors (KNN) Logistic Regression Naive Bayes	The dataset was extremely imbalanced and highly skewed. The following sampling techniques were performed: RUS, ROS, SMOTE and Adasyn (Adaptive Synthetic). The Logistic Regression achieved the highest accuracy (99,94%) after under sampling techniques and Random Forest achieved the highest accuracy (99,96%) after over sampling techniques. Out of the four sampling techniques, SMOTE performed very well.

Reference	Title	Fraud Type	Methods	Findings
(Kotsiantis, Koumanakos, Tzelepis & Tampakas, 2007)	Forecasting Fraudulent Financial Statements using Data Mining	Financial Statement	Decision Trees Neural Networks Bayesian Network SVM Nearest Neighbor Ensemble	The best model in terms of accuracy was the Ensemble classifier, using a stacking variant methodology, having an accuracy of 95,1%. Decision Trees also achieved a good performance, having an accuracy of 91,2%.
(Liu, 2008)	Fraudulent financial reporting detection and business failure prediction models: A comparison	Financial Statement	Neural Networks Logistic Regression Decision Trees	Two models were implemented: prediction of business failure and prediction of fraudulent financial reporting. The Logistic Regression was the one with the best performance in both models having an accuracy of 99,05% on predicting business failure and 96,5% of accuracy in predicting fraudulent financial reporting.
(Deng & Mei, 2009)	Combining self-organizing map and k-means clustering for detecting fraudulent financial statements	Financial Statement	Self-Organization Map (SOM) K-means Clustering	An unsupervised learning approach was used for the prediction. The authors designed a clustering model called V-KSOM (combination of Self-Organization Map (SOM) and K-means clustering).
(Viaene, Dedene & Derrig, 2005)	Auto claim fraud detection using Bayesian learning neural networks	Insurance Fraud	Neural Networks Logistic regression Decision Trees	Neural Networks results were compared to the Logistic regression and Decision Trees classifiers, to find the best model.
(Yang & Hwang, 2006)	A process-mining framework for the detection of healthcare fraud and abuse	Insurance Fraud	Structure Pattern Algorithm Classification Based on Associations Algorithm (CBA)	Structure pattern algorithm was used to find patterns in the data. The predictive model (CBA) was created using the features of the structure pattern discovery algorithm as inputs. The detection model proved to be efficient and capable of identifying fraudulent cases.
(Thiprungsri & Vasarhelyi, 2011)	Cluster analysis for anomaly detection in accounting data: An audit approach	Insurance Fraud	K-means Clustering	Using K-means clustering procedure, eight clusters were formed. The authors concluded that cluster analysis is a useful audit technology and a good option for fraud and anomaly detection.
(Rodrigues & Omar, 2014)	Auto Claim Fraud Detection Using Multi Classifier System	Insurance Fraud	Decision Trees Naïve Bayes Neural Networks SVM Ensemble	Working with an imbalanced dataset. To solve this problem, subsample generation was used. The champion model was the combination of all models – the Ensemble method, using an average vote function decision.
(Hassan & Abraham, 2016)	Modeling Insurance Fraud Detection Using Imbalanced Data Classification	Insurance Fraud	Decision Trees SVM Neural Networks	Imbalanced data distribution. Two methods were applied: under-sampling, with replacement and without-replacement. Decision Trees where the algorithm that performed better.
(Tang & Yin, 2005)	Developing an intelligent data discriminating system of anti-money laundering based on SVM	ML	Statistical Learning Theory (SLT) SVM	Statistical Learning Theory (SLT) was used to improve the embarrassments of anti-money laundering. SVM algorithm was used to detect unusual behaviour.
(Keyan & Tingting, 2011)	An improved support-vector network model for anti-money laundering. Management of e-Commerce and eGovernment (ICMeCG)	ML	SVM	Find the optimal SVM classifier parameters using the cross-validation method, based on the highest classification accuracy rate.
(Lv, Ji & Zhang, 2008)	An RBF neural network model for anti-money laundering	ML	Neural Networks	Radial Basis Function (RBF) Neural Network model to track ML activity. The model could reach high correction rate, reducing false positive rate and enhancing positive rate.
(Le Khac & Kechadi, 2010)	Application of data mining for anti-money laundering detection: A case study	ML	K-means Clustering Neural Networks Genetics Algorithm Heuristic Algorithm	The authors concluded that their approach could satisfy the needs of detecting money laundering. The implemented solution can improve the performance of the current CE bank's solution in terms of running time.
(Liu, Qian, Mao & Zhu, 2011)	Research on anti-money laundering based on core decision tree algorithm	ML	Decision Trees BIRCH Algorithm K-means Clustering	Cluster analysis (combining the BIRCH and K-means algorithm) was used to identify typical money laundering patterns and rules. Decision Tree algorithm was applied to detect abnormal transaction data.
(Khalaf Ahmed Allam El-Din & El Khamesy, 2016)	Data Mining Techniques for Anti-Money Laundering	ML	Neural Networks	Four types of neural networks were tested: Multi-Layer Perceptron Neural Network (MLP), Probabilistic Neural Network (PNN), Radial Basis Function (RBF) and Linear Neural Network (LNN).

Reference	Title	Fraud Type	Methods	Findings
				The Linear Neural Network was the model with best performance, having 80% of data correctly classified.
(Sain & Puri, 2018)	Detection of money laundering accounts using data mining techniques	ML	n.a.	Based on a research review conducted with an emphasis on data mining techniques for fraud detection, the authors presented different methodologies for detecting suspicious accounts involved in ML activities.
(Lokanan, 2022)	Predicting Money Laundering Using Machine Learning and Artificial Neural Networks Algorithms in Banks	ML	Naive Bayes Logistic Regression Random Forest Catboost Algorithm Artificial Neural Networks	A total of 5 machine learning algorithms were trained and tested using data that came from a simulator of money-laundering activities in Middle Eastern banks based on a real dataset. The Naive Bayes and Random Forest models were the two best-performing models to predict money laundering transactions, having both an accuracy equal to 77,5%.

Summarizing, several studies over the past years have shown data mining techniques and the application of different algorithms to predict financial fraud (e.g. money laundering, credit card fraud, financial statement fraud and insurance fraud). The available studies tend to be higher for credit card fraud when compared to money laundering. These prior studies demonstrate the growing interest in data mining and machine learning to detect and predict fraud behaviour or illegal activities.

The supervised algorithms most used over the years are the following: Neural Networks, Logistic Regression, Support Vector Machine, Naive Bayesian, Decision Trees and Random Forest. Recent studies show the usage of Boosting algorithms to predict fraudulent transactions. Each algorithm was able to show its advantages, achieving high-accuracy results and an overall good performance. Nonetheless, the performance of each model varied depending on the type of fraud being predicted, on the approach presented by the author, if the dataset had problems of class imbalance and in the resampling technique used to solve the imbalance problem.

4. METHODOLOGY

Data mining has experienced significant growth and widespread adoption among academics, analysts, and scientists over the past few years. Data mining is used to identify data patterns and extract valuable knowledge. A suitable technique must be employed to extract the desirable knowledge. Otherwise, there is the risk of losing valuable insights and essential information contained in the dataset (Chen et al., 1996) (Simoudis, 1996) (Fayyad, 1996).

In this study, the Exploratory Data Analysis (EDA) and the predictive models will be created using R software. The EDA approach employs a variety of techniques that help (Yu, 2010):

- Maximize the knowledge of the dataset;
- Uncover underlying patterns in the data;
- Identify important variables;
- Detect outliers and anomalies in the data;
- Test underlying assumptions;
- Develop and implement the predictive models; and,
- Determine optimal factor settings.

Two types of machine learning algorithms are used in fraud detection: supervised and unsupervised. Supervised learning uses already classified data – labelled as fraud activity, for example – to learn complex patterns. Unsupervised learning deals with datasets with no label and infers inner data structure. The dataset used for this study has a priori transactions classified as (1) 'fraud', identifying money laundering transactions and (2) 'no fraud', representing legal transactions.

Since the dataset is highly imbalanced, the class to study fraud has a lower event rate, being the minority class. Imbalanced class distribution is expected in anomaly detection scenarios like electricity pilferage, fraudulent banking transactions, identification of rare diseases, etc. If this problem is not treated, the developed predictive models using conventional machine learning algorithms could be biased and inaccurate (Weh & Yusuf, 2019). This happens because machine learning algorithms are usually designed to improve accuracy by reducing errors. Thus, they do not consider the class distribution.

The solution to the imbalanced dataset problem is to resample the training dataset. The most well-known resampling methods are Random-Under Sampling (RUS), Random-Over Sampling (ROS), Cluster-Based Over Sampling, Synthetic Minority Over-Sampling Techniques (SMOTE) and Algorithmic Ensemble Techniques. The most used resampling techniques are RUS, ROS and SMOTE (Chawla et al., 2002) (Wang & Yao, 2009) (Nguyen et al., 2018) (Bhattacharyya et al., 2011) (Hordri et al., 2018) (Rajora et al., 2018).

For this study, the chosen solution combines the K-means clustering algorithm with SMOTE. This technique proved effective in overcoming the imbalance between and within classes and avoiding the generation of noise. It outperformed popular oversampling methods (Douzas, Bação & Last, 2018).

The following supervised learning algorithms will be applied to predict money laundering transactions: Logistic Regression, Neural Networks, Decision Trees, Random Forests, Light Gradient Boost and Ensemble.

The performance of the models will be evaluated considering the confusion matrix results and the following measures: accuracy, precision, F1-score, Gini coefficient and AUC score.

The aim is to choose the model with the best performance. The model should present a high accuracy, precision and detection rate. It is also essential to consider a high value for F1-score, which ensures a balance between precision and recall since we are dealing with an imbalanced dataset and a high AUC score that shows the model's discriminatory power.

In the following figure, it is presented the methodology model that illustrates the process followed in this dissertation:

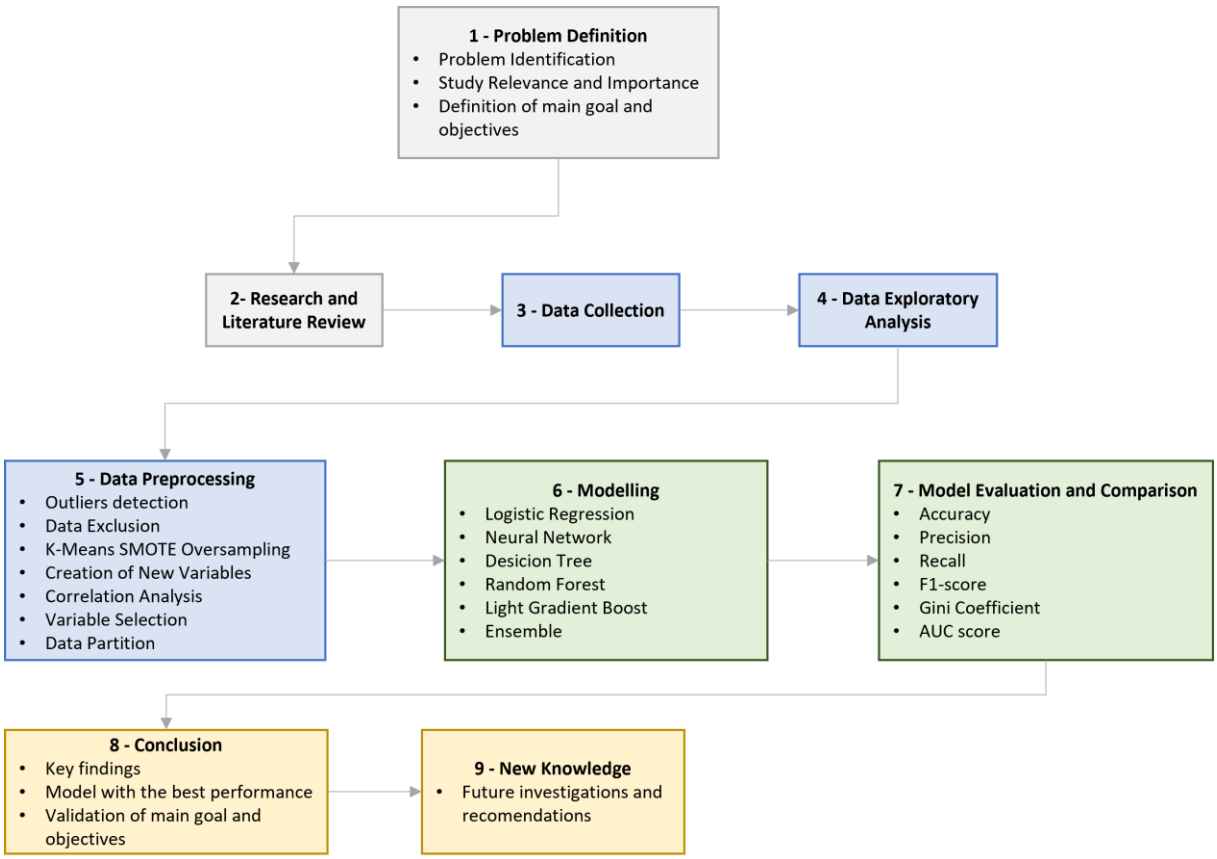

Figure 1 - Dissertation Methodology

4.1. Sample Data

The dataset used in this study is a public dataset from Kaggle[1]. This synthetic dataset is generated using a PaySim simulator that is scaled down to ¼ of the original dataset.

PaySim simulates mobile money transactions based on a sample of real-life transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The

[1] https://www.kaggle.com/

E. A. Lopez-Rojas, A. Elmir, and S. Axelsson. "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016

objective of PaySim is to use this data to simulate legal operations and transactions and to inject suspicious behavior to evaluate the performance of fraud detection models.

The dataset has 11 variables and a total of 6.362.620 bank transactions. Each transaction is classified as considered fraud or a legal transaction. The table below presents the description of all variables.

Table 2 - Dataset Variables Description

Variable Name	Description	Variable Type
Step	Maps a unit of time in the real world. One step is 1h of our time. Total steps has 743, making 30 days of simulation.	Numeric
Type	Type of the transaction.	Character
Amount	Amount of the transaction in local currency ($ - dollar).	Numeric
NameOrig	Customer who started the transaction (remitter).	Character
OldbalanceOrig	Initial balance before the transaction.	Numeric
NewbalanceOrig	New balance after the transaction.	Numeric
NameDest	Customer who is the receiver of the transaction (beneficiary).	Character
OldbalanceDest	Initial balance of the beneficiary.	Numeric
NewbalanceDest	New balance of the beneficiary.	Numeric
IsFraud	Flag that indicates if the transactions are fraudulent or normal transactions.	Numeric
IsFlaggedFraud	Flag that indicates illegal attempts considered to be massive transfers from one account to another. An illegal attempt is a single transaction with an amount over than 200.000.	Numeric

This study's target variable (dependent variable) is the variable “isFraud”. In this context, it identifies fraudulent transactions carried out by individuals within the simulation. More specifically, it refers to fraudulent activity made by the agents when they try to illicitly take control of the client accounts and launder the money by transferring it to another account, after which the funds are then withdrawn. This binary variable identifies money laundering transactions when the value is equal to 1 (isFraud = 1) and non-fraudulent transactions when the value is equal to 0 (isFraud = 0).

4.2. DATA EXPLORATION

The PaySim dataset has a total of 6.362.620 bank transactions over 30 days. There are no missing values in the dataset.

The number of transactions identified as fraudulent is 8.213, corresponding to 0,1% of the total transactions. The number of non-fraudulent transactions is 6.354.407, which is 99.8% of the total transactions.

Table 3 - Frequency of Transactions by Target Variable

Target	Number of Transactions	Percentage (%)
1	8.213	0,13
0	6.354.407	99,87

This shows we are dealing with a severe class imbalance since the event rate for fraudulent transactions is less than 1%. To solve the imbalance problem, the K-Means SMOTE technique will be applied to increase the fraud cases in the dataset synthetically. This sampling technique is covered in the following chapter, Data Preprocessing.

As illustrated in the figure below, there are five unique transaction types: Cash In, Cash Out, Debit, Payment and Transfer. “Cash Out” is the most frequent transaction type, corresponding to 35% of the total transactions, followed by Payment with 34%. The transaction type “Cash In” accounts for 22% of all transactions, while the least frequent transaction types are Transfer with 8% and Debit with approximately 1%.

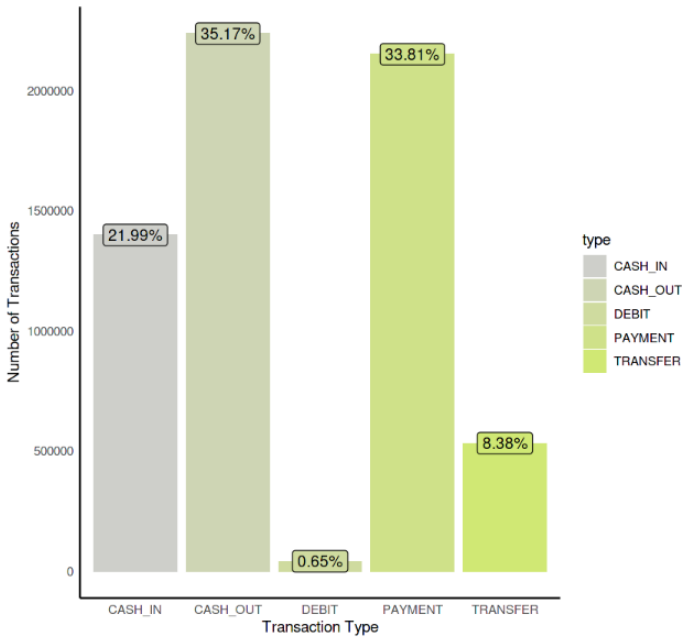

Figure 2 - Distribution of Transactions by Transaction Type

Fraudulent transactions only occur for two types of transactions: Cash Out and Transfer. The remaining transaction types of Cash In, Debit and Payment are never fraud-related. As depicted in the table below, out of the 2.237.500 Cash Out transactions, about 4.116 are classified as fraud. Regarding Transfer type, from the 532.909 total transactions, 4.097 are identified as fraudulent.

Table 4 - Frequency of Transaction Type by Target Variable

Type	n	Fraud (1)	No Fraud (0)
CASH_IN	1.399.284	0	1.399.284
CASH_OUT	2.237.500	4.116	2.233.384
DEBIT	41.432	0	41.432
PAYMENT	2.151.495	0	2.151.495
TRANSFER	532.909	4.097	528.812

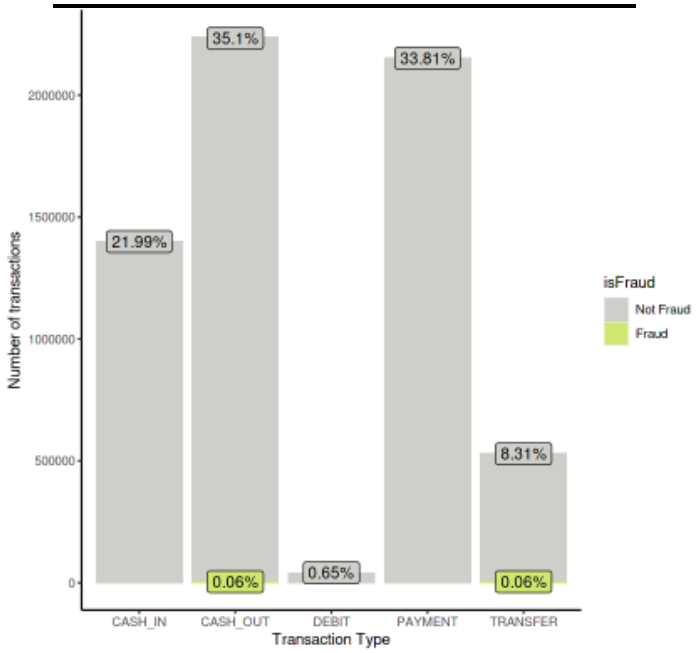

Figure 3 - Distribution of Transaction Type by Target Variable

Analyzing the descriptive statistics of the numeric variables in the table below, we can conclude that the average "Amount" of money transferred is 179.862$. The highest money transferred is around 92$ million, while the minimum value is set to zero. Analyzing the remaining minimum values of all the "balance" variables, they also present a minimum value equal to zero. Since we are dealing with transactional data, these transactions amounting to zero don't seem correct, at least for two variables – transaction amount ("Amount") and new balance of the beneficiary ("NewbalanceOrig"). Since a simulator generated the dataset, we will assume these zero amounts are an error caused by the simulator system.

The maximum value before the transfer ("OldbalanceOrig") and the maximum value after the transfer ("NewbalanceOrig") are very similar, having a difference of 10$ million. The average initial balance of the beneficiary ("OldbalanceDest") is around 1$ million, while the average new balance of the beneficiary ("NewbalanceDest") shows a higher value at 124.294$. It is worth noting that the difference between the maximum value before the transfer ("OldbalanceOrig") and the balance of the beneficiary after the transfer ("NewbalanceDest") was over 269$ million. When analyzing the average for these variables, the difference is higher at 391.113$.

Table 5 - Statistics of Numeric Variables

Numeric Variable	Mean	Std. Deviation	Median	Min	Max
Step	243,4	142,332	239	1	743
Amount	179.862	603.858,2	74.872	0	92.445.517
OldbalanceOrig	833.883	2.888.243	14.208	0	59.585.040
NewbalanceOrig	855.114	2.924.049	0	0	49.585.040
OldbalanceDest	1.100.702	3.399.180	132.706	0	35.6015.889
NewbalanceDest	1.224.996	3.674.129	214.661	0	356.179.279

By analysing the average results and the standard deviation for the "Amount" variable and all the "balance" variables, it can also be concluded that the observations are widely spread out from the mean. The maximum values of these variables suggest that significantly higher values in the dataset contribute to this large standard deviation results. The high standard deviation and large maximum values indicate that we are dealing with a wide range of values and are strong evidence of outliers in the dataset.

Focusing only on the money laundering transactions, the average amount of money laundered is around 1$ million, corresponding to the exact amount of 1.467.967$, and the highest amount laundered is precisely 10$ million. These results show that large amounts of money are used to perform illegal transfers between accounts and that money laundering constitutes a serious and significant problem for the banking sector.

Round numbers in transaction amounts can be a red flag for fraudulent activity, as they are linked to unique transactions where exact amounts would be more common. This round effect was studied by Nigrini (2016), where the investigation of round numbers uncovered financial fraud in accounting textbooks and test banks. In this dataset, round amounts occur more frequently in fraudulent transactions than in standard transactions, although it covers only 6% of the money laundering transactions.

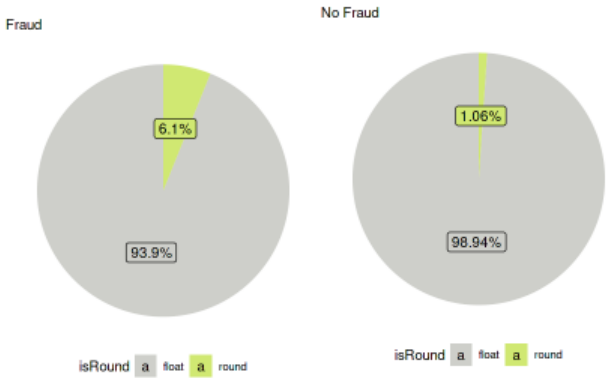

Figure 4 - Identification of Round Amounts by Target Variable

Analyzing the distribution of transactions over time, for the 30 days, illustrated in the figure below, the non-fraudulent transactions show a uniform distribution with some fluctuations depending on the

hours that the transaction was performed since non-fraudulent transactions tend to occur more frequently during business hours. But overall, there is a consistent number of transactions over different hours. In the middle of the month, there is a drop in the frequency of non-fraudulent transactions, which indicates a period of lower transactional activity. This behaviour is aligned with a study by CaixaBank in 2021, which shows that the spending percentage is higher during the first week of the month and tends to decrease towards the end of the month, resulting in lower transaction activity. Focusing on fraudulent transactions, the distribution over time shows a more variable pattern than non-fraudulent transactions. It's observable for money laundering transactions and consecutive periods of higher and lower activity. Although the number of fraudulent transactions varies more significantly over time, it's important to note that money laundering transactions occur every hour of the timeline. It's undeniable that money laundering activity is present every hour of the month.

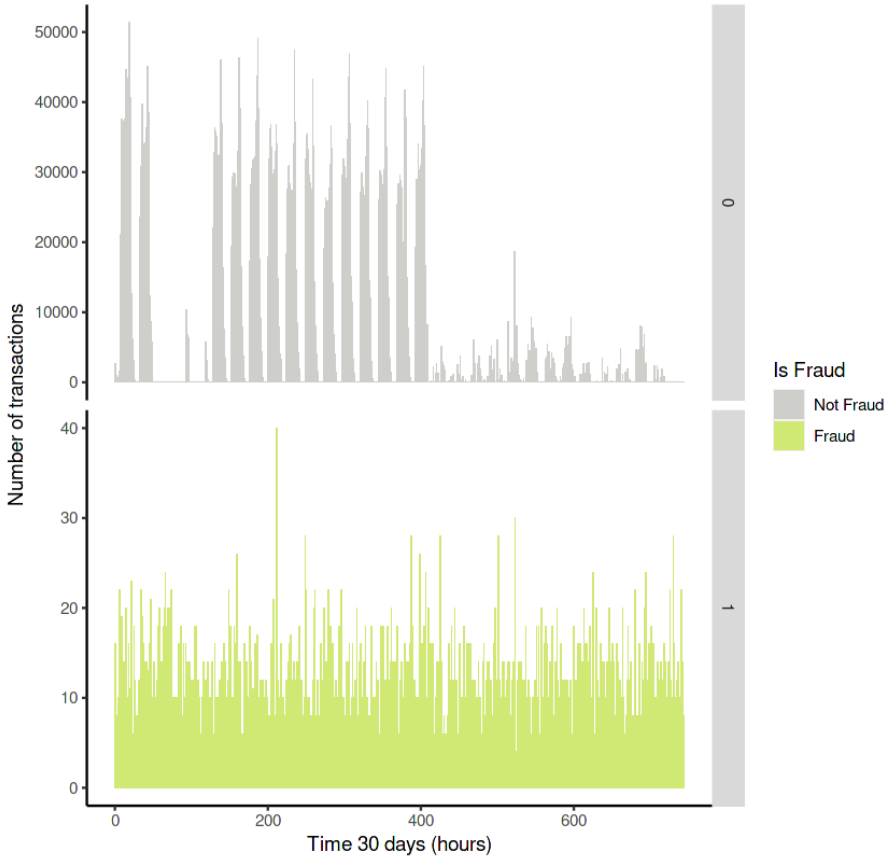

Figure 5 - Distribution of Transactions over Time (30 days)

In this dataset, besides the target variable, there is another binary variable named "IsFlaggedFraud". This variable was created by the simulator that generated the dataset based on a business rule developed to detect cases of fraudulent transactions. This business rule aims to detect illegal attempts of money transfers in a single transaction, considering only the currency amount as a condition to flag the transaction. If the amount exceeds 200.000$, the transaction will be flagged as fraud ("IsFlaggedFraud" =1). Analyzing this variable, there are only 16 transactions flagged as fraudulent, which is a minimal number since the dataset has a total of 1.673.570 transactions with an amount higher than 200.000$. With this finding, we can conclude that the simulator was not precise in flagging the transactions and that this variable does not provide important information for this study. As a result, it was decided to exclude this variable from the modelling exercise.

Regarding the categorical variables, “NameOrig” and “NameDest”, these variables are responsible for identifying the customer who started the transaction (“NameOrig”) and the beneficiary who received the transaction (“NameDest”). It’s an alphanumeric code where each identification number starts with a ‘C’ or ‘M’ prefix. “NameOrig” only has identification numbers that begin with the ‘C’ prefix, whereas “NameDest” includes identification numbers with both prefixes. After analyzing these variables and confronting them with the target variable, we can conclude that fraudulent transactions occur only for individuals identified with the ‘C’ prefix. Individuals identified with the ‘M’ prefix only associate with non-fraudulent transactions.

These findings provide a clear understanding of the dataset’s variables, characteristics and underlying patterns of the data, allowing us to proceed 4with the preprocessing phase to prepare the data for the predictive models.

4.3. Data Preprocessing

This phase involves preparing the data for the modelling exercise by correcting the gaps identified during exploratory data analysis. This chapter will detail the preprocessing steps undertaken, including handling outliers imbalanced data, transforming and creating new variables to enhance the model’s predictive power, encoding categorical variables, correlation analysis, variable selection and splitting data into training and validation sets.

4.3.1. Outliers

Over the years, many authors have proposed varying definitions for outliers or anomalies. Still, the most widely recognized definition is by Hawkins, who defines the concept of outliers as follows: “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that a different mechanism generated it” (Hawkins, 1980) (Hilal, Gadsden & Yawney, 2022). In this sense, outliers can be described as extreme values that stand out significantly from the overall pattern of values in the dataset.

Outliers can have a significant impact on the predictive model estimates and also on statistical analysis, potentially causing biased results. This leads to the fundamental question of whether outliers should be or not removed. However, eliminating outliers is legitimate only for specific reasons and circumstances. According to Hitt et al., outliers can also be of substantive interest and studied as unique phenomena that may lead to novel theoretical insights (Hitt et al., 1998).

During the data exploration phase, we identified that some numeric variables could have outliers. The presence of outliers in the dataset can be justified by the fact that we are dealing with transactions (e.g. day-to-day regular transactions, business transactions and fraudulent transactions), capturing a wide range of activities and behaviours.

Considering the abovementioned points, we decided to keep the outliers in the dataset. This decision ensured that the models did not disregard valuable information from the outlier observations.

4.3.2. Data Exclusion

As identified during the data exploration phase, some cases are never classified as money laundering activity, as follows:

- Fraudulent transactions only occur for Cash Out and Transfer types. The remaining transaction Types – Cash In, Debit and Payment, are never associated with money laundering transactions, and,
- Fraudulent transactions coincide only for individuals identified with the 'C' prefix regarding the "NameDest" variable. Records identified with the 'M' prefix are always non-fraudulent transactions.

As an approach to have a dataset more representative of the positive class ("isFraud" = 1), we decided to exclude the records identified in both situations listed above (where "isFraud" = 0). This decision was made with the objective to improve the predictive model's ability to identify cases of money laundering activity, thus improving their performance and predictive accuracy in detecting fraudulent transactions.

It was then removed from the dataset a total of 3.592.211 non-fraudulent records. This is equivalent to 56.46% of the data. The following table shows the number of transactions and percentage, by fraudulent and non-fraudulent transactions after the removal of records.

Table 6 - Frequency of Transactions by Target variable after Data Exclusion

Target	Number of Transactions	Percentage (%)
1	8.213	0,29
0	2.762.196	99,7

The dataset now contains a total of 2.770.409 transactions. As it can be observed, the number of non-fraudulent transactions is 2.762.196, and there has been a slight adjustment in the percentage of fraudulent transactions, which corresponds now to approximately 0,3%.

The event rate for fraudulent transactions remains extremely low. The dataset continues to be highly imbalanced. Therefore, the next chapter will focus on addressing the imbalance problem by implementing the oversampling technique K-Means SMOTE.

4.3.3. K-Means SMOTE Oversampling

The dataset used in this study is highly imbalanced. This happens when the classes of data are not equally represented. In this case, we are dealing with banking transactions classified as fraudulent or non-fraudulent. By the nature of the business, it's natural to exhibit a minority class corresponding, in this case, to the fraudulent transactions.

The dataset started with an event rate for fraudulent transactions equal to 0,1%, less than 1%, after removing some non-fraudulent transactions, explained in detail in the previous chapter 4.3.2. Data Exclusion: the event rate suffered a minor adjustment, becoming approximately 0,3%.

Such a severe imbalance needs to be treated, otherwise it will impact the predictive model's performance and classification accuracy, resulting in biased results. This can be solved by applying an oversampling technique, adding instances to the minority class using duplication or generation of new samples (Kotsiantis et al., 2006).

The chosen oversampling technique for this study combines the K-means clustering algorithm with SMOTE, a method proposed by Douzas et al. (2018). This method consists of three steps: clustering, filtering and oversampling—the clustering step groups data using the K-means algorithm. The filter step chooses clusters to be oversampled and determines how many samples need to be generated for each cluster. The oversampling step is when the oversampling using SMOTE is applied, but only in clusters dominated by the minority class. This technique proved to be effective in overcoming the imbalance between and within classes and avoiding noise generation by oversampling only in safe areas. It outperformed popular oversampling methods (Douzas, Bação & Last, 2018).

To surpass the imbalance problem, the K-means SMOTE method was then applied to the dataset, resulting in the synthetic creation of 1.234.775 fraudulent transactions. The table below shows the updated total transactions number by fraudulent transactions.

Table 7 - Frequency of Transactions by Target variable after K-means SMOTE

Target	Number of Transactions	Percentage (%)
1	1.242.988	31,04
0	2.762.196	68,96

The proportion of fraudulent transactions (minority class) is now 31%. According to Khan et al. (2023), if the proportion of the minority class is between 20-40% of the dataset, the degree of imbalance is considered mild. The classification of degree imbalance is illustrated in the figure below.

Degree of Data Imbalance	Proportion of Minority Class (%)
Severe	<1% of the dataset
Moderate	1–20% of the dataset
Mild	20–40% of the dataset

Figure 6 - Classification of Degree Imbalance for Imbalanced Data (Khan et. al, 2023)

The severe imbalance problem of dataset is corrected after applying the oversampling technique. Although the classes are not evenly distributed, the dataset can be used for modelling since the degree of data imbalance is mild.

4.3.4. Creation of New Variables

New variables were created to get more insight into the data and improve the models' performance and predictive power. Some variables alone may not be considered relevant, but when combined with other variables or when appropriately treated, they can prove to be significant for the models. The following table shows a summary of the new variables created.

Table 8 - New Variables Definition

Variable Name	Description	Variable Type
iszero_oldbalanceOrig	Identifies if oldbalanceOrig has an amount equal to zero or not.	Numeric

Variable Name	Description	Variable Type
iszero_oldbalanceDest	Identifies if oldbalanceDest has an amount equal to zero or not.	Numeric
iszero_newbalanceOrig	Identifies if newbalanceOrig has an amount equal to zero or not.	Numeric
iszero_newbalanceDest	Identifies if newbalanceDest has an amount equal to zero or not.	Numeric
balance_of_origin_dif	Difference between newbalanceOrig and oldbalanceOrig variables.	Numeric
balance_of_dest_dif	Difference between newbalanceDest and oldbalanceDest variables.	Numeric
orig_balance_error	Identification of balance errors for the “Orig” variables. Formula: oldbalanceOrig - newbalanceOrig - amount.	Numeric
dest_balance_error	Identification of balance errors for the “Dest” variables. Formula: newbalanceDest - oldbalanceDest - amount.	Numeric
is_round	Identifies if the amount is a round or float number.	Numeric
hour	Identifies the time of the day during 24h.	Numeric

In this step, the categorical variable Type was transformed into a dummy variable that has a value equal to 1 if the transaction type is “Transfer” or equal to 0 if the transaction type is “Cash out”.

4.3.5. Correlation

Correlation informs us about the degree of association between variables. It evaluates the strength and direction of the relationship between two or more variables. Correlation coefficient is the measure to quantify the degree of relationship of the variables, with values ranging from -1 to 1. The interpretation of the correlation coefficient corresponds to the following:

- A positive correlation coefficient means that when one variable changes, either up or down, the other variable changes in the same direction;
- A negative correlation coefficient means that when one variable changes, the other variable changes in the opposite direction, and,
- A zero-correlation coefficient means that there is no linear relationship between the variables.

The correlation matrix was drawn to analyse and better understand the relationships and dependencies of the variables. This analysis aimed to identify highly correlated variables that need to be excluded from the modelling exercise to avoid multicollinearity problems. In this study, we considered a high correlation when the correlation coefficient is equal to or greater than 0,8.

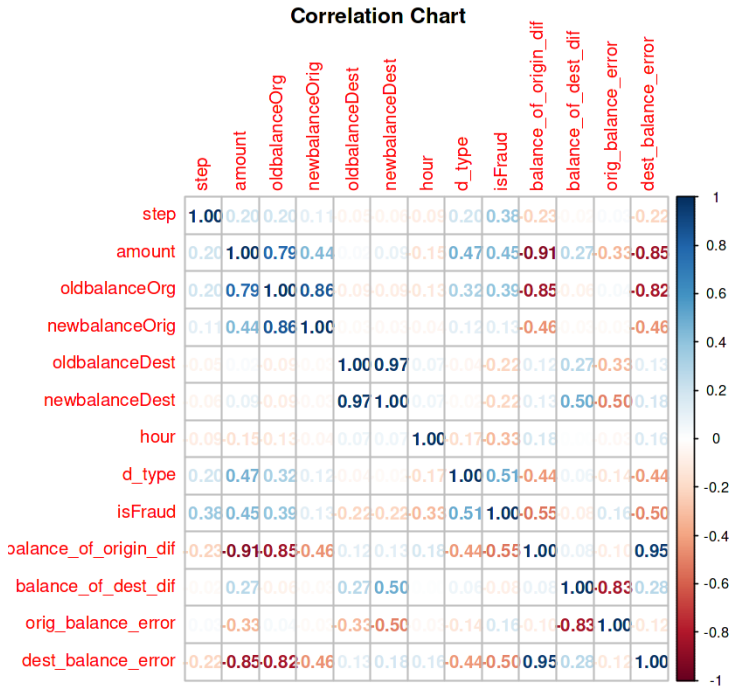

Figure 7 - Correlation Matrix

Analysing the correlation matrix, we can observe the following high correlations between variables:

- Variable “dest_balance_error” is highly correlated with three other variables. It shows a strong relationship with “Amount” (correlation coefficient = -0,85), with “oldbalanceOrig” (correlation coefficient = -0,82) and with “balance_of_origin_dif” (correlation coefficient = 0,95);
- Variable “amount” is also highly correlated with “balance_of_origin_dif” (correlation coefficient = -0,91);
- Variable “balance_of_origin_dif” is also highly correlated with “oldbalanceOrig” (correlation coefficient = 0,95);
- Variable “oldbalanceOrig” is also highly correlated with “newbalanceOrig” (correlation coefficient = 0,86);
- Variable “oldbalanceDest” is highly correlated with “newbalanceDest” (correlation coefficient = 0,97); and,
- Variable “orig_balance_error” is highly correlated with “balance_of_dest_dif” (correlation coefficient = -0,83).

Summarizing, the following variables present a strong linear relationship with other variables of the dataset: “dest_balance_error”, “amount”, “balance_of_origin_dif”,; “oldbalanceOrig”; “newbalanceDest”, and “orig_balance_error”.

4.3.6. Variable Selection

During the data exploration phase, it was identified that the “IsFlaggedFraud” variable is not precise in identifying transactions with a currency amount higher than 200.000$ and that the categorical variables “NameOrig” and “NameDest” are alphanumeric codes that start with ‘C’ or ‘M’ prefix. These

variables do not seem to provide important information for this study, so it was decided that they should be excluded from the modelling exercise.

The variables “iszero_oldbalanceOrig”, “iszero_oldbalanceDest”, “iszero_newbalanceOrig”, and “iszero_newbalanceDest” present zero variance thus they will not be selected for the modelling exercise.

Taking in consideration the correlation results for the numeric variables presented in the previous chapter, the following variables were excluded from the modelling exercise: “oldbalanceOrg”, “oldbalanceDest”, “balance_of_origin_dif”, “orig_balance_error” and “dest_balance_error”.

The table below lists the selected variables to be the input variables for training the predictive models.

Table 9 - Variable Selection for the Predictive Models

Variable	Role	Variable Type
IsFraud	Target	Numeric
Step	Input	Numeric
Amount	Input	Numeric
NewbalanceOrig	Input	Numeric
NewbalanceDest	Input	Numeric
Hour	Input	Numeric
D_type	Input	Numeric
Balance_of_dest_dif	Input	Numeric
IsRound	Input	Numeric
IsFlaggedFraud	Rejected	Numeric
NameOrig	Rejected	Character
NameDest	Rejected	Character
Iszero_oldbalanceOrig	Rejected	Numeric
Iszero_oldbalanceDest	Rejected	Numeric
Iszero_newbalanceOrig	Rejected	Numeric
Iszero_newbalanceDest	Rejected	Numeric
OldbalanceOrg	Rejected	Numeric
OldbalanceDest	Rejected	Numeric
Balance_of_origin_dif	Rejected	Numeric
Orig_balance_error	Rejected	Numeric
Dest_balance_error	Rejected	Numeric

4.3.7. Data Partition

Before advancing to the model phase, the dataset was split into two sets: the training set and the validation set. The training set is used to train and develop the models, whereas the validation set is used to validate the performance of the models. This splitting ensures that the models are accurate and is essentially done to avoid overfitting (Khan, 2014).

The dataset was split into 70% of the records for the training set and the remaining 30% for the validation set. This splitting approach allowed the allocation of more observations to train the models.

The training set has a total of 2.803.629 transactions, where 1.933.917 are non-fraudulent, and 869.712 are fraudulent.

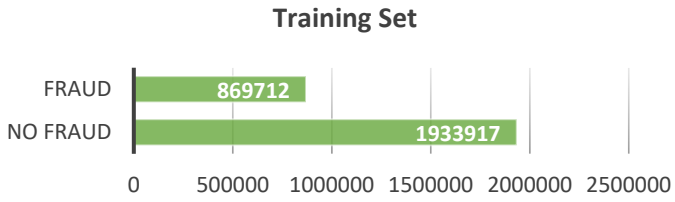

Figure 8 - Training Set

While the validation set has a total of 1.201.555 transactions, where about 828.279 are non-fraudulent transactions and 373.276 are fraudulent transactions.

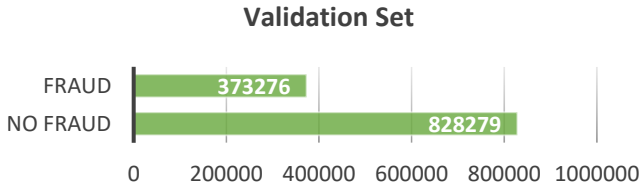

Figure 9 - Validation Set

4.4. MODELLING

The model phase consists of developing and training the predictive models with the selected input variables from the dataset. As identified in the literature review, different machine learning algorithms are used to predict situations of fraud behaviour, such as logistic regression, neural networks, SVM, random forest, ensemble, and boosting algorithms. Each algorithm as advantages and performance varied depending on the type of fraud it was being predicated on (e.g., Credit Card fraud, Financial Statement fraud, Insurance fraud or Money Laundering), depending on the approach presented and depending as well on the data mining situation (e.g., data sources and problems of imbalanced data).

This study decided to implement the following supervised learning algorithms to predict money laundering transactions: Logistic Regression, Neural Networks, Decision Trees, Random Forest, Light Gradient Boost (LGB) and Ensemble.

4.4.1. Logistic Regression

Logistic Regression (LR) continues to be one of the most important data mining techniques used for classification problems. It predicts a binary outcome based on a set of independent variables. The binary outcome is set only in two possible classes – 1 (yes) and 0 (no) (Maalouf, 2011).

The output of the LR model is a probability between 0 and 1, representing the likelihood of the event happening. LR is based on the Maximum Likelihood estimation algorithm, which means that the regression coefficients should be selected to maximize the probability of the Y (dependent variable) given X (independent variable) (Bhalla, 2014).

As discussed, the LR model is one of the most used models to predict fraud behaviour. It proved that it can achieve a high level of accuracy when predicting business failure (accuracy = 99,05%) and when predicting fraudulent financial reporting (accuracy = 96,5%) (Liu, 2008). This result led to the decision to include the LR model in the scope of this study.

The LR model was trained using the training set. Based on what was learned from the training, the model was executed with the validation set to predict the labels of the target variable. The tables below show the confusion matrix for the training and validation set.

Table 10 - Confusion Matrix of Logistic Regression (Training Set)

		Predicted	
		0	1
Actual	0	1.894.412	39.505
	1	95.462	774.250

Table 11 - Confusion Matrix of Logistic Regression (Validation Set)

		Predicted	
		0	1
Actual	0	811.216	17.063
	1	40.928	332.348

The LR model performance is analyzed thoroughly in the results and discussion chapter.

4.4.2. Neural Networks

Over the years, Neural Networks (NN) have been applied to a variety of domains, different expertise areas, and different business problems, and they have shown widespread usage.

NN can be defined as a computational model based on the structure and functions of a biological neural network. The NN architecture consists of three layers of nodes or artificial neurons: the input, hidden, and output layers. Each node is connected, allowing them to communicate with each other, and each connection has its associated weight. The input layer receives the data that will be used to train the model, the hidden layer passes the information obtained from the input layer multiplied by adjusted weights, and the output layer predicts the final estimate (1 or 0) (Qamar et al, 2023).

As discussed previously, the NN model is one of the most used models to predict cases of fraud behavior. It has been used for Credit Card fraud, Financial Statement fraud, Insurance fraud and for Money Laundering prediction. Regarding this last business problem, Money laundering, NN proved that it can achieve a high accuracy when predicting money laundering transactions (accuracy=80%) (Khalaf Ahmed Allam El-Din & El Khamesy, 2016). This result led to the decision to include the NN model in the scope of this study.

The NN model was trained with the training set. Based on what was learned from the training, the model was executed with the validation set to predict the labels of the target variable. The tables below show the confusion matrix for the training and validation set.

Table 12 - Confusion Matrix of Neural Network (Training Set)

		Predicted	
		0	1
Actual	0	1.929.248	4.669
	1	249.661	620.051

Table 13 - Confusion Matrix of Neural Network (Validation Set)

		Predicted	
		0	1
Actual	0	826.254	2.025
	1	107.079	266.197

The NN model performance is analyzed thoroughly in the results and discussion chapter.

4.4.3. Decision Trees

A Decision Tree (DT) is a non-parametric supervised machine learning algorithm. It can be described as a decision support tool that uses a tree-like model of decisions and their possible consequences. The fact that DT is a non-parametric model means that it has no assumption about the data type or other characteristics of the variables, such as distribution, outliers or missing values. The main objective is to develop a strategy to maximize the accuracy of the prediction (Breslow & Aha, 1997).

DT is a hierarchical collection of rules that describe how to divide the dataset into smaller groups. The process involves deciding the characteristics and conditions (i.e. rules) to apply to splitting and knowing when splitting should stop (Gupta, 2017).

As discussed previously, the DT model is one of the most used models to predict cases of fraud behavior. It has been used for Credit Card fraud, Financial Statement fraud, Insurance fraud and for Money Laundering prediction. DT proved that it can achieve a high accuracy when predicting financial statement fraud (accuracy = 91,2%). This result led to the decision to include the DT model in the scope of this study.

The DT model was trained with the training set. Based on what was learned from the training, the model was executed with the validation set to predict the labels of the target variable. The tables below show the confusion matrix for the training and validation set.

Table 14 - Confusion Matrix of Decision Tree (Validation Set)

		Predicted	
		0	1
Actual	0	1.893.290	40.627
	1	92.699	777.013

Table 15 - Confusion Matrix of Decision Tree (Validation Set)

		Predicted	
		0	1
Actual	0	810.894	17.385
	1	39.595	333.681

The DT model performance is analyzed thoroughly in the results and discussion chapter.

4.4.4. Random Forest

The Random Forest (RF) model consists of multiple decision trees. RF can be described as an ensemble learning algorithm of several decision trees. In other words, the RF creates a “forest” built from several Decision Trees, which are trained using bagging methods. It outputs the mean of their prediction to correct the individual Decision Trees' tendency to overfit the data. It can be used for multiple regression and classification problems (Cutler et al., 2011).

As discussed previously, the RF model is one of the most used models to predict cases of fraud behavior. It has been used for credit card fraud and money laundering prediction. DT proved that they can achieve a high accuracy when predicting Credit Card fraud after over sampling techniques, having an accuracy equal to 99,96% (Bhuiyan et al., 2022). It also achieved an accuracy of 77,5% when predicting Money Laundering (Lokanan, 2022). This result led to the decision to include the RF model in the scope of this study.

The RF model was trained using the training set. Based on what was learned from the training, the model was executed with the validation set to predict the labels of the target variable. The tables below show the confusion matrix for the training and validation set.

Table 16 - Confusion Matrix of Random Forest (Training Set)

		Predicted	
		0	1
Actual	0	1.881.919	51.998
	1	2.076	867.636

Table 17 - Confusion Matrix of Random Forest (Validation Set)

		Predicted	
		0	1
Actual	0	805062	23217
	1	1425	371851

The RF model performance is analyzed thoroughly in the results and discussion chapter.

4.4.5. Boosting Algorithm

Boosting is a powerful ensemble technique that has shown success in various practical applications and in different business areas. It combines the predictions of multiple weak learners to create a single, more accurate strong learner. The most popular weak learner is the Decision Tree due to their ability to work with any dataset. The main idea of boosting is to add new models to the ensemble sequentially, where each new model is trained to minimize the loss function, such as the mean squared error of the previous model. The new model's predictions are then added to the ensemble, and the process is repeated until a stopping criterion is met (Natekin et al., 2013).

As discussed previously, Boosting Algorithms have been used to predict cases of Credit Card fraud. Boosting algorithms proved that they can achieve a high level of prediction accuracy, where the best model of the study conducted by Gamini et al. was the Catboost algorithm with an accuracy equal to 92% (Gamini et al., 2021). This result led to the decision to include the Light Gradient Boost (LGB) model in the scope of this study. The light method was chosen since it was designed to handle large datasets, being more efficient and outperforming in training speed compared to the Extreme Gradient Boost.

The LGB model was trained using the training set. Based on what was learned from the training, the model was executed with the validation set to predict the labels of the target variable. The tables below show the confusion matrix for the training and validation set.

Table 18 - Confusion Matrix of Light Gradient Boost (Training Set)

		Predicted	
		0	1
Actual	0	1.920.254	13.663
	1	31.363	838.349

Table 19 - Confusion Matrix of Light Gradient Boost (Validation Set)

		Predicted	
		0	1
Actual	0	822.344	5.935
	1	13.541	359.735

The LGB model performance is analyzed thoroughly in the results and discussion chapter.

4.4.6. Stacking Ensemble

The Ensemble can be described as combining multiple models to obtain a more robust model than its constituents. The Ensemble focuses on improving the multiple tested individual algorithms by combining two or more trained models to get a more accurate model than any individual model. An Ensemble learning can reduce the risk of overfitting due to the diversity of the combined models (Mohammed & Kora, 2023).

Over the years, the Ensemble algorithm has been applied successfully in several domains. As discussed previously, Ensemble has been used to predict fraud behaviour, such as credit card fraud, financial statement fraud, and insurance fraud. It proved that it can achieve high accuracy, for example, when predicting Financial Statement fraud, with an accuracy equal to 95,1% (Kotsiantis et al., 2007). This result led to the decision to include the Ensemble model in the scope of this study.

The stacking ensemble model was trained with the training set using the following trained models: logistic regression, neural networks, decision trees, and random forest. Based on what was learned from the training, the model was executed with the validation set to predict the labels of the target variable. The tables below, show the confusion matrix for the training and validation set.

Table 20 - Confusion Matrix of Ensemble (Training Set)

		Predicted	
		0	1
Actual	0	1.881.889	52.028
	1	2.102	867.610

Table 21 - Confusion Matrix of Ensemble (Validation Set)

		Predicted	
		0	1
Actual	0	805.061	23.218
	1	1.434	371.842

The Ensemble model performance is analyzed thoroughly in the results and discussion chapter.

4.5. Model Evaluation

The performance of the models is evaluated by considering the confusion matrix results and the following measures: accuracy, precision, recall, F1-score, Gini coefficient, and AUC measure.

The confusion matrix contains information about the actual and predicted classifications, shown in a matrix format as the name imply. The columns represent the predicted values, while rows represent the actual values, or vice versa (Fig. 9).

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Figure 10 - Confusion Matrix

In the context of this study, each entry of the confusion matrix can be interpreted as (Modi, 2017):

- True Positive (TP) - number of correctly classified fraudulent transactions;
- True Negative (TN) - number of correctly classified non-fraudulent transactions;
- False Positive (FP) - number of incorrectly classified non-fraudulent transactions; and,
- False Negative (FN) - number of incorrectly classified fraudulent transactions.

From the confusion matrix, the following metrics can be derived (Modi, 2017) (Last, 2017) (Japkowicz, 2013):

- Accuracy: is the ratio of total number of correct predictions and the total number of predictions. Is a measure of how well the model performs. The accuracy can be measured on a scale of 0 to 1 or as a percentage. Using the accuracy measure it's possible to obtain its inverse, the error rate.

$$Accuracy\ (A) = \frac{(TP+TN)}{(TP+FN+TN+FP)};\ \text{Error Rate}\ =\ 1 -\ \text{Accuracy} \tag{1}$$

- Precision: also known as positive predictive value. Is a metric that shows how often the model correctly predicts the positive class. The precision can be measured on a scale of 0 to 1 or as a percentage. A high precision means a low false positive rate and indicates that the model is predicting the target class correctly.

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{2}$$

- Recall: also known as sensitivity or true positive rate. Shows how well a model can find all objects of the target class. The recall can be measured on a scale of 0 to 1 or as a percentage. A high recall means that the model is capturing nearly all positives.

$$Recall\ (R) = \frac{TP}{(TP + FN)} \tag{3}$$

- F1-score: is the harmonic mean of precision and recall. The indicator rates both the completeness and exactness of positive predictions. The values of F1-score varies between 0 and 1. A higher F1-score indicates a good model performance.

$$F1 - score = 2 * (\frac{Precision * Recall}{Precision + Recall}) \tag{4}$$

The accuracy is a measure that is impacted by imbalanced datasets, showing a bias toward the majority class (Last, 2017). Since the dataset in this study was treated with the K-means SMOTE oversampling technique to become more balanced, it was decided to keep this measure for the model assessment. However, this measure is analysed with precaution for each model.

Regarding AUC measure, it stands for Area Under the ROC curve and is a metric used to assess the overall model performance. The ROC is a graph that maps the true positive rate (TPR) and the false positive rate (FPR). The AUC score corresponds to this area under the curve, meaning that the score represents the ability of the model to predict classes correctly. The AUC values range between 0 and 1. A high value for AUC means that the model is very likely to assign a higher predicted probability to a random positive instance (fraudulent transactions) than to a random negative one (non-fraudulent transactions). This means the model has strong discriminatory power and better predictive performance (Bradley, 1997).

In what concerns Gini coefficient, this is a metric that indicates the model discriminatory power. Meaning, the capacity and effectiveness of the model in differentiating between the positive class and the negative class. The Gini coefficient ranges between 0 and 1, where 0 represents perfect equality (same as saying no discrimination) and 1 represents perfect inequality (meaning perfect discrimination) (Bendel et al, 1989). In the context of modelling, a higher Gini coefficient indicates better model performance in terms of its ability to accurately rank transactions based on fraudulent or non-fraudulent.

5. RESULTS AND DISCUSSION

This study aims to create and train a model that is able to predict fraudulent transactions, where the type of fraud in scope is money laundering. A total of six predictive models were trained and tested using different algorithms: Logistic Regression, Neural Network, Decision Tree, Random Forest, Light Gradient Boost and Ensemble.

During the modelling exercise, it was identified that the variables with the most importance to the model predictions were the following: new balance of the beneficiary ("NewbalanceDest"); beneficiary balance difference ("Balance_of_dest_dif"); currency amount ("Amount") and transaction type ("Type").

To assess the performance of each model, the select metrics for model evaluation were computed for the validation dataset, being illustrated on Table 22.

Table 22 - Model Performance Results

	Logistic Regression	Neural Network	Decision Tree	Random Forest	LGB	Ensemble
Accuracy	0,952	0,909	0,953	0,979	0,984	0,979
Error Rate	0,048	0,091	0,047	0,021	0,016	0,021
Precision	0,951	0,992	0,950	0,941	0,984	0,941
Recall	0,890	0,713	0,894	0,996	0,964	0,996
AUC	0,988	0,963	0,936	0,984	0,999	0,984
Gini Coefficient	0,976	0,925	0,873	0,968	0,998	0,968
F1-score	0,920	0,830	0,921	0,968	0,974	0,968

Table 22 shows that for each model, the accuracy and precision have similar or close results, even equal values for the Light Gradient Boost, which indicates that each model is correctly identifying a high number of fraudulent transactions and suggests that the number of false positives is low among all predictions of each model.

Logistic Regression performs similarly to the Decision Tree model, where the most visible differences are in the AUC score and Gini Coefficient. The Logistic Regression achieved a higher AUC score (98,8%) and a higher Gini Coefficient (0,976) than the Decision Tree model.

Neural Network was the model that achieved the highest precision, with 99,2%. However, the model registered the lowest recall (71,3%) and consequently the lowest f1-score (83%). Regarding prediction accuracy, it is also the model with the lowest accuracy (90,9%).

Random Forests and Ensemble demonstrate an equal performance for the selected evaluation metrics. These two models registered the highest recall (99,6%) among all models. The precision is a little lower compared to the other models, having a value equal to 94,1%.

Light Gradient Boost is the model with the highest accuracy (98,4%), AUC score (99,9%), Gini Coefficient (0,998) and f1-score (97,4%). In terms of precision and recall, is the second lead model having a precision equal to 98,4%, which differs from the Neural Network in 0,8%, and a recall equal

to 96,4%, which differs from the Random Forest and Ensemble in 3,2%. It is also the model with the lowest error rate (1,6%) among all the models.

Considering these results, it can be concluded that the Light Gradient Boost is the model with the best performance, showing a strong discriminatory power. It also shows a strong precision (98,4%) and recall (96,4%). The highest value of the f1-score (97,4%) reveals the balance between precision and recall, showing that the model correctly identifies a high number of fraudulent transactions while minimizing the false positives and false negatives.

6. CONCLUSION

This dissertation focused on one specific type of financial fraud - Money Laundering, which consists of taking cash earned from illegal activities and making the cash appear to be gained from a legal activity (Hilal et al., 2022).

Money Laundering is a severe threat to financial institutions. According to the United Nations Office on Drug and Crime, the global value of laundered money in one year ranges between 500 billion dollars to 1 trillion dollars (Le Khac & Kechadi, 2010). Considering this, it is important for a business to prevent and detect fraud behaviour in real time in order to avoid money losses, fines from the regulator and to avoid the exposure to financial and operational risk.

The main goal of this study was to create a predictive model in the Financial Industry, more specifically in the Banking sector, to detect suspicious cases of money laundering using transactional data. The proposed approach was the creation of six different supervised learning algorithms to predict fraudulent transactions, being the following: Logistic Regression, Neural Networks, Decision Trees, Random Forest, Light Gradient Boost and Ensemble.

Since the dataset used in this study was highly imbalanced, it was necessary to apply an oversampling technique. The chosen solution was the combination of K-means clustering algorithm with SMOTE. This technique proved effective in overcoming the imbalance between and within classes and avoiding noise generation (Douzas, Bação & Last, 2018).

Regarding the influence of the variables, it can be concluded that the variables with the most importance to the model predictions were the following: new balance of the beneficiary ("NewbalanceDest"); beneficiary balance difference ("Balance_of_dest_dif"); currency amount ("Amount") and transaction type ("Type").

The Light Gradient Boost was the model with the best performance, showing a strong discriminatory power. It also shows a strong precision (98,4%) and recall (96,4%). The highest value of f1-score (97,4%) among all the models reveal the balance between precision and recall, showing that the model is correctly identifying a high number of fraudulent transactions, while minimizing the false positives and false negatives.

When comparing to similar studies, Gamini et al. in 2021 used only Boosting algorithms to predict credit card fraudulent transactions. The Catboost algorithm was the best model having an accuracy equal to 92%, lower than our model by 6,4%, and a recall equal to 1, which in this case is higher than our model by 3,6%. Nonetheless, the recall obtained in our model is considered to be high showing that our model has a strong capacity in identifying correctly the majority of fraudulent transactions. This proves the effectiveness of our model in accurately identifying the positive class.

In conclusion, this dissertation proved that data mining techniques can continuously be used to detect cases of fraud behaviour, especially cases of financial fraud in the Banking sector. In the available literature the focus of many authors has been mainly directed to predict credit card fraud, having few studies available related to money laundering detection. This study proved that by monitoring and analysing transaction data, fraudulent transactions can be predicted with high levels of success

achieved. This study has the power to influence and improve current implemented money laundering detection processes in the banking sector.

This dissertation also intends to be a positive contribute to academics and researchers in the area of data mining and financial fraud, aiming to detect and prevent cases of money laundering.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORK

During the development of this dissertation, the principal limitation found was the lack of a feasible public dataset related to money laundering transactions. After intense research, we were able to find a public synthetic dataset generated using a simulator called PaySim. If real data was used in the scope of this study, the findings and results obtained could be confirmed in practice.

The fact that the dataset is generated from a simulator, may lead to some data quality problems that were identified during the data exploration phase. Since the dataset is also based in mobile payments, the dataset has a small number of features. It contains only the necessary information to perform a mobile payment.

Another limitation of the dataset is that the proportion of fraudulent transactions was severely low, being necessary to apply an oversampling technique to solve the problem of imbalanced data. Lastly, the dataset had only available transaction information that occur during one month.

As a recommendation for future work, it is proposed to apply the models developed in this dissertation using a real dataset with more balance between classes (i.e. fraudulent transactions and non-fraudulent transactions) and that covers a longer period of transaction activity, for example, six months to one year of transactional data. With a bigger time-interval, it would be possible to create a transactional profile for each account. It would also be beneficial if the information regarding the remitter country and the beneficiary country of the transaction was available. From that, it would be possible to classify each country with a risk level (e.g. high, medium, low) based on financial security and safety, to identify if the country is affected by terrorism and lasty, if the country is an offshore. It would also be interesting if the dataset had information related with the account (e.g. account type and bank product) and related with the first holder of the account (e.g. identification if it is an individual or organization, age, economic activity, country of residence, country of citizenship, identification if it is a political exposed person (PEP), identification if it is a sanctioned or non-sanctioned, identification if it is relative or close associate to PEP, etc.).

Another suggestion is to apply other supervised algorithms to detect money laundering, such as SVM, Naive Bayes or apply other boosting algorithms, and compare their performance with the results obtained in this study.

Last recommendation for future work, would be to apply the models considered in this dissertation to other types of financial fraud to evaluate their performance and effectiveness, such as Credit Card fraud, Financial Statement fraud and Insurance Fraud. It can also be explored the usage of this models in less research areas of fraud, such as securities and commodities fraud, mortgage fraud, insider trading and others.

8. BIBLIOGRAPHICAL REFERENCES

Albashrawi, M., & Lowell, M. (2016). Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. Journal of Data Science, 14(3), 553–570.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conference Data Mining, (January), 182–185.

Bendel, R. B., Higgins, S. S., Teberg, J. E., & Pyke, D. A. (1989). Comparison of skewness coefficient, coefficient of variation, and Gini coefficient as inequality measures within populations. Oecologia, 78(3), 394–400. https://doi.org/10.1007/BF00379115

Bhalla, D. (2014). Difference between Linear Regression and Logistic Regression. Retrieved from https://www.listendata.com/2014/11/difference-between-linear-regression.html

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

Bhuiyan, R. A., Khatun, M. S., Taslim, M., & Hossain, M. A. (2022). Handling Class Imbalance in Credit Card Fraud Using Various Sampling Techniques. American Journal of Multidisciplinary Research and Innovation (AJMRI), 1(4), 160. ISSN: 2158-8155 (Online), 2832-4854.

Bradley, Andrew P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. Vol. 30, Issue 7, Pages 1145-1159, ISSN 0031-3203. https://doi.org/10.1016/S0031-3203(96)00142-2

Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural Data Mining for Credit Card Fraud Detection. In IEEE International Conference on Tools with Artificial Intelligence ICTAI-99 (pp. 103–106).

Breslow, L. A., & Aha, D. W. (1997). Simplifying decision trees: A survey. Knowledge Engineering Review. https://doi.org/10.1017/S0269888997000015

CaixaBank (2021) How do we spend throughout the month? How do we spend throughout the month? (caixabankresearch.com)

Chawla, N., & Bowyer, K., & Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357/jair.953 https://doi.org/10.1613/jair.953

Chen, M., Han, J., & Yu, P.S. (1996). Data Mining: An Overview from a Database Perspective. IEEE Trans. Knowl. Data Eng., 8, 866-883.

Cutler, A., Cutler, D.R., Stevens, J.R. (2012). Random Forests. In: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9326-7_5

Deng, Q., & Mei, G. (2009). Combining self-organizing map and k-means clustering for detecting fraudulent financial statements. In 2009 IEEE International Conference on Granular Computing, GRC 2009. https://doi.org/10.1109/GRC.2009.5255148

Douzas, G., Bação, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, 1-20. https://doi.org/10.1016/j.ins.2018.06.056

Dharwa, J., & Patel, A. (2011). A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction. International Journal of Computer Applications. 16. 10.5120/1977-2651.

Egiyi, M., & Chindengwike, J. D. (2023). The Psychology Behind Financial Fraud: Unmasking Motives and Warning Signs. 4. 16-24. 10.5281/zenodo.8287913.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In Fayyad, U. M.et al (Eds.), Advances in knowledge discovery and data mining. AAAI Press / The MIT Press.

Gamini, P., & Yerramsetti, S., & Darapu, G., & Pentakoti, V., & Prudhvi, V., & Professor, Assistant, & Student, & Communication (2021). Detection of Credit Card Fraudulent Transactions using Boosting Algorithms.

Gupta, P. (2017). Decision Trees in Machine Learning. Retrieved from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Hassan, A., & Abraham, A. (2016). Modeling Insurance Fraud Detection Using Imbalanced Data Classification. 10.1007/978-3-319-27400-3_11.

Hawkins, D.M. (1980). Identification of Outliers. Monographs on Applied Probability and Statistics, Vol. 11. London: Chapman and Hall.

Hitt, M.A., Harrison, J.S., Ireland, R.D., & Best, A. (1998). Attributes of Successful and Unsuccessful Acquisitions of US Firms. British Journal of Management, 9, 91-114.

Hordri, N., & Sophiayati, S., & Firdaus, N., & Mariyam, S. (2018). Handling Class Imbalance in Credit Card Fraud using Resampling Methods. International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.091155 http://dx.doi.org/10.14569/IJACSA.2018.091155

Japkowicz, N. (2013). Assessment metrics for imbalanced learning. In He, H. and Ma, Y., editors, Imbalanced learning, pages 187–206. John Wiley & Sons.

Keyan, L., & Yu, T. (2011). An Improved Support-Vector Network Model for Anti-Money Laundering. 2011 Fifth International Conference on Management of e-Commerce and e-Government, 193-196.

Khalaf, A., & El, N. (2016). Data Mining Techniques for Anti-Money Laundering. International Journal of Computer Applications, 146, 28-33.

Khan, A., & Malim, N. (2023). The classification of the degree of imbalance for imbalanced data. Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-classification-of-the-degree-of-imbalance-for-imbalanced-data_tbl1_368817166

Khan, R. Z. (2014). Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (comparative study). Computer Science, Communication & Instrumentation Devices, (December 2014), 163–172.

Khan, A., & Malim, N. (2023). Comparative Studies on Resampling Techniques in Machine Learning and Deep Learning Models for Drug-Target Interaction Prediction. Molecules (Basel, Switzerland). 28. 10.3390/molecules28041663.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Handling imbalanced datasets: A review. Science, 30(1):25–36

Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2007). Forecasting Fraudulent Financial Statements using Data Mining. Proceedings of World Academy of Science, Engineering and Technology, vol 12.

Last, F., Douzas, G., & Bação, F. (2017). Oversampling for Imbalanced Learning Based on K-Means and SMOTE.

Le Khac, N. A., & Kechadi, M. T. (2010). Application of data mining for anti-money laundering detection: A case study. In Proceedings - IEEE International Conference on Data Mining, ICDM. https://doi.org/10.1109/ICDMW.2010.66

Liou, F.M. (2008). Fraudulent financial reporting detection and business failure prediction models: A comparison. Managerial Auditing Journal. https://doi.org/10.1108/02686900810890625

Liu, R., Qian, X., Mao, S., & Zhu, S. (2011). Research on anti-money laundering based on core decision tree algorithm. Proceedings of the 2011 Chinese Control and Decision Conference, CCDC 2011. 10.1109/CCDC.2011.5968986.

Lokanan, M. (2022). Predicting Money Laundering Using Machine Learning and Artificial Neural Networks Algorithms in Banks. Journal of Applied Security Research. 19. 1-25. 10.1080/19361610.2022.2114744.

Lo, S., & Li, T. (2016). Using Big Data Analytics for Money Laundering Detection – A Case Study.

Lopez-Rojas, E.A., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus.

Lv, L., Ji, N., & Zhang, J. (2008). An RBF neural network model for anti-money laundering. 1. 209 - 215. 10.1109/ICWAPR.2008.4635778.

Maalouf, M. (2011). Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335.

Mills, C. (2017). Predictive Analytics in fraud and AML. Journal of Financial Compliance, Vol. 1, No. 1, 17-26.

Modi, K. (2017). Fraud Detection Technique in Credit Card Transactions using Convolutional Neural Network. International Journal of Advance Research in Engineering, Science & Technology. 4. 2394-2444.

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University - Computer and Information Sciences, Vol. 35, Issue 2, 757-774. ISSN 1319-1578. https://www.sciencedirect.com/science/article/pii/S1319157823000228

Mukherjee, S., Mukherjee, T., & Nath, A. (2016). Fraud Analytics Using Data Mining. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 3(4), 1–11.

Natekin, A., & Knoll, A. (2013). Gradient Boosting Machines, A Tutorial. Frontiers in neurorobotics. 7. 21. 10.3389/fnbot.2013.00021.

Nigrini, M. (2016). The Implications of the Similarity Between Fraud Numbers and the Numbers in Financial Accounting Textbooks and Test Banks. Journal of Forensic Accounting Research. 1. 10.2308/jfar-51465.

Pimenta, C. (2009). Esboço de Quantificação da Fraude em Portugal (1st ed.). Edições Húmus. ISBN 978-989-8139-08-5. http://www.gestaodefraude.eu

Pinzón, N., Koundinya, V., Galt, R., Dowling, W., Boukloh, M., Taku-Forchu, N.C., Schohr, T., Roche, L., Ikendi, S., Cooper, M.H. and Parker, L.E. (2023). AI-Powered Fraud and the Erosion of Online Survey Integrity: An Analysis of 31 Fraud Detection Strategies (No. 95tka). Center for Open Science.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. Monash University.

PwC (2018) Global Economic Crime and Fraud Survey. PwC's 2018 Global Economic Crime and Fraud Survey | PwC

PwC (2020) Global Economic Crime and Fraud Survey. PwC's Global Economic Crime and Fraud Survey 2020

Qamar, R., & Zardari, B. (2023). Artificial Neural Networks: An Overview. Mesopotamian Journal of Computer Science. 2023. 130-139. 10.58496/MJCSC/2023/015.

Rajora, S., Li, D., Jha, C., Bharill, N., Patel, O.P., Joshi, S., Puthal, D., & Prasad, M. (2018). A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance. 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 1958-1963.

Rodrigues, L. A., & Omar, N. (2014). Auto Claim Fraud Detection Using Multi Classifier System, 37–44. https://doi.org/10.5121/csit.2014.4604

Sain, P., & Puri, S. (2018). Detection of money laundering accounts using data mining techniques.

Salehi, A., Ghazanfari, M., & Fathian, M. (2017). Data Mining Techniques for Anti Money Laundering. International Journal of Applied Engineering Research ISSN 0973-4562, Vol. 12, No. 20, 10084-10094.

Simoudis, E. (1996). Reality check for data mining. IEEE Explore, Vol. 11, No. 5, 26-33.

Tang, J., & Yin, J. (2005). Developing an intelligent data discriminating system of anti-money laundering based on SVM. 2005 International Conference on Machine Learning and Cybernetics, 6, 3453-3457 Vol. 6.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. International Journal of Digital Accounting Research. https://doi.org/10.4192/1577-8517-v11_4

Viaene, S., Dedene, G., & Derrig, R. A. (2005). Auto claim fraud detection using Bayesian learning neural networks. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2005.04.030

Yang, W. S., & Hwang, S. Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2005.09.003

Yeh, I. C., & Lien, C. hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2007.12.020

Yu, C.H. (2010). Exploratory data analysis in the context of data mining and resampling.

Hilal, W., & Gadsden, S. A., & Yawney, J. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. Expert Systems with Applications. Volume 193. 116429. ISSN 0957-4174 https://doi.org/10.1016/j.eswa.2021.116429

Wang, S., & Yao, X. (2009). Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models. 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings. 324-331. 10.1109/CIDM.2009.4938667.

Wei, W., Li, J., Cao, L., Ou, Y., & Chen, J. (2012). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16, 449 - 475.

Wen, S. W & Yusuf, R. M. (2019). Predicting Credit Card Fraud on an Imbalanced Data. International Journal of Data Science and Advanced Analytics, Vol 1, 12-17.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. Computers and Security. https://doi.org/10.1016/j.cose.2015.09.005

West, J., & Bhattacharya, M. (2016). An investigation on experimental issues in financial fraud mining. In Proceedings of the 2016 IEEE 11th Conference on Industrial Electronics and Applications, ICIEA 2016. https://doi.org/10.1109/ICIEA.2016.7603878