NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

USING PREDICTIVE ANALYTICS FOR MONEY MULE DETECTION ON A CRYPTOCURRENCY EXCHANGE

Katja Kerzic

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

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

BOOK SPINE


2022
Using Predictive Analytics for Money Mule Detection on a Cryptocurrency Exchange
Katja Kerzic
MGI

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

USING PREDICTIVE ANALYTICS FOR MONEY MULE DETECTION ON A CRYPTOCURRENCY EXCHANGE

by

Katja Kerzic

Dissertation presented as a partial requirement for obtaining the Master's degree in Information Management, with a specialisation in Information Systems and Technologies Management

Advisor / Co-Advisor: Ian Scott, PhD

Co-Advisor*:* Luka Tomat, PhD

November 2022

ACKNOWLEDGEMENTS

This dissertation is part of a double degree program, an opportunity I greatly appreciate, as it allowed me to spend each year of my Master's studies at a different university.

I would like to thank my supervisors, Professor Ian Scott, PhD from NOVA IMS Information Management School and Professor Luka Tomat, PhD from the University of Ljubljana, School of Economics and Business, for their expertise and support throughout the writing process.

This dissertation topic would not have been possible without the cryptocurrency exchange's cooperation and several employees' technical support.

I would also like to thank my parents for their support and encouraging statements like "Are you still not done with the dissertation?" and "When was the last time you saw the sun?" after spending days, weeks, and months behind a computer screen. Most importantly, I thank my siblings for always giving me new unreachable achievements to follow.

ABSTRACT

Cryptocurrencies and blockchain, the novel technology that became widespread with Bitcoin implementation in 2009, offer many applications. While the new technology facilitates fast and pseudo-anonymous transactions, it also leaves room to be exploited for illicit activities, such as money laundering. This dissertation focuses on developing a predictive analytics model with supervised machine learning algorithms for classifying money mule fraud instances.

Money mules are an instrument for the layering stage of money laundering and are used by criminals to hide the origins of their wealth. As confirmed cases were a rare event, the algorithms used were optimised for imbalanced data in addition to trying resampling techniques for under and oversampling the dataset. The algorithms used were logistic regression, decision trees and ensemble methods – random forest and two types of gradient boosting: XGBoost and LightGBM. The most promising results were achieved with the random forest algorithm, as it reached the best result metrics values aligned with the given business objective. However, the study concluded that the business objective could not be fully realised with the developed model, as it falsely predicts a percentage of the events, which could cause constraints on the business. Therefore, the selected company can use the results and models as a reference tool and a base for further data analyses.

KEYWORDS

Cryptocurrency; Money Laundering; Money Mule; Predictive Analytics; Machine Learning

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AML	Anti-Money Laundering
AUC	Area Under the Receiver Operating Characteristic Curve
BTC	Bitcoin
CA	Classification Accuracy
CRISP-DM	Cross-Industry Standard for Data Mining
DLT	Distributed Ledger Technology
DT	Decision Tree
ETH	Ethereum
EUROPOL	The European Union Agency for Law Enforcement Cooperation
FATF	Financial Action Task Force
FBI	Federal Bureau of Investigation
GBM	Gradient Boosting Models
GBP	British Pound Sterling
GDP	Gross Domestic Product
KYC	Know Your Customer
LGBM	Light Gradient Boosting Machine (LightGBM)
LR	Logistic Regression
MR	Misclassification Rate
PoW	Proof of Work
RF	Random Forest
USD	United States Dollar
XGB	eXtreme Gradient Boosting (XGBoost)
XRP	Ripple

1. INTRODUCTION

Money mules are used to hide the origin of money gained through illicit activity. They are an instrument for money laundering used in the placement and layering of the money laundering process. The money mule's role is to transfer the money, thus helping hide the origin, which usually earns them compensation. Using mules to hide the origin of assets is a practice that has been used throughout history; however, with the increasing use of digital currencies, the practice moved from more traditional financial institutions like banks to cryptocurrency exchanges and other services (Esoimeme, 2021; EUROPOL, 2021). Working for a cryptocurrency exchange (selected company) provided insight into issues surrounding money mules, which created an opportunity to further research and develop a dissertation on the topic.

The selected company is a centralised exchange and needs to adhere to strict compliance regulations and anti-money laundering (AML) practices. Therefore, it is essential to minimise illicit activities on the platform, preferably using preventative measures (Arslanian and Fischer, 2019). This dissertation does not negate the potential of decentralisation, nor does it focus on the role of exchanges in the crypto space. The emphasis will be on analysing data collected on a cryptocurrency exchange in a specific recent period and how that data could be used to predict money mule fraud occurrences. The approach to achieving the objective is predictive analytics, specifically classification using supervised machine learning algorithms.

The purpose of the dissertation is to develop a predictive analytics model that will enable the selected company to detect more money mules and minimise the risk of illicit user behaviour. The research and data analysis will follow the Cross-Industry Standard for Data Mining (CRISP-DM) methodology, which offers a structured approach to different types of data analyses. The methodology consists of six stages, starting with an overview of the business understanding of the issue, which is integral for developing a predictive model that will be aligned with business goals. The goal is to develop predictive models using different supervised machine learning algorithms. The algorithms used for the model development are logistic regression, decision tree, and three ensemble methods. These models will then be evaluated based on several classification metrics. This process will ensure the selection of the most appropriate final model, which should be aligned with the purpose of the dissertation while also reaching the business objectives. The data analysis process and results will be comprehensively presented and discussed while complying with data security standards concerning the selected company.

The dissertation begins with a thorough literature review to better understand the cryptocurrency sector and money laundering. That will be followed by general predictive analytics and specific fraud detection applications. Sources used in the dissertation are both primary and secondary, whereas the critical literature review includes recognised scientific and professional literature collected through various internet databases.

2. LITERATURE REVIEW

The dissertation topic consists of several different areas, which is why chapter 2, Literature Review, comprises information gathered from secondary sources on the main topics of the Master's dissertation. The goal is to present the theoretical framework of these main topics, including cryptocurrencies, fraud, and predictive analytics and its use for fraud detection.

The literature review starts with a general overview of blockchain and cryptocurrencies to gain insight into their emergence. These subchapters will be followed by insight into relevant fraud occurrences, followed by, more specifically, money mules and their role in financial fraud. They are followed by information about fraud in the cryptocurrency industry and how it can affect different market participants. After the basic topics are covered, the focus will be on how predictive analytics can be used to classify fraudulent behaviour.

2.1. Blockchain and Cryptocurrency

The emergence of cryptocurrency and blockchain in popular use dates back to 2008, which was also the year of one of the most recent global financial crises. More precisely, January 2009 is when the Bitcoin network started operation. It all started when Satoshi Nakamoto published a "white paper" that presented Bitcoin as "the purely peer-to-peer version of electronic cash that would allow online payments to be sent directly from one party to another without going through a financial institution" (Nakamoto 2008). Firstly, let us clarify that Bitcoin and blockchain are not synonymous. The emergence of the terms being used 2interchangeably is related to the start of Bitcoin use, as it runs on a blockchain, along with the majority of other cryptocurrencies. However, even though Bitcoin runs on a blockchain, the word blockchain was not mentioned in the Bitcoin white paper but merely as a "chain of blocks" and only came into widespread use in 2015 after its use in some popular news articles. Blockchain is one of the distributed ledger technologies (DLT) but not the only one. Some cryptocurrencies like HBAR run on a DLT that is not blockchain. HBAR is the native cryptocurrency of a patented DLT called Hedera Hashgraph. There are numerous ways to use blockchain technology, and its use for cryptocurrencies is just one. Other possibilities include various smart contracts, supply chain management, healthcare, network security and many more (Arslanian and Fischer 2019; Lu 2018).

To emphasise that cryptocurrencies are not the only possible application, a more inclusive expression is "crypto assets" (Arslanian and Fischer, 2019). In this dissertation, the terms are used interchangeably.

Blockchain is a technology that provides secure, transparent, and efficient digitally signed transaction chains. It is a decentralised peer-to-peer technology that facilitates integrity using cryptography. Similar to the conventional public ledger, the blocks in the blockchain keep all the transaction records. These blocks are connected into a chain that grows with each new block. The technology has many differentiating characteristics, but some of the key ones are (Zheng et al., 2017):

- Decentralisation – transactions are carried out without the need to be confirmed by a centralised third party (for example, a bank) but are supported by consensus algorithms.
- Persistency – fast validation of transactions, invalid transactions, and the blocks in which they are held can be discovered immediately.

- Anonymity – an individual's interaction with the blockchain can theoretically be anonymous (more in subchapter 2.1.2. Transactions on the Blockchain).
- Auditability – on the Bitcoin blockchain, a new transaction is carried out through previous unspent transactions, which means each transaction can be easily tracked.

There are three types of blockchain technologies, depending on their accessibility, presented in Table 2.1. The blockchain can be public, hybrid, or private (Lu, 2018).

Table 2.1 – Types of blockchain depending on accessibility

Type of blockchain	How it works	Applications
Public	Fully decentralised blockchain. Anyone can access data on nodes.	Bitcoin and other public cryptocurrencies.
Consortium (hybrid)	Partially distributed (multi-centre) blockchain. Predefined nodes control the consensus process.	Hyperledger – enterprise application solution
Private	Centralised blockchain for internal organisation management.	Healthcare and supply chain management

Source: Adapted from Lu (2018)

2.1.1. Cryptocurrencies

The idea behind cryptocurrency creation was to be used as a financial instrument and offer an alternative to the centralised financial system. Cryptocurrencies are digital currencies that use cryptography to ensure transaction security. The word crypto is derived from cryptography and relates to secure communication. Cryptocurrencies are currencies that use secure transactions based on proof, and the blockchain on which they run uses an asymmetric cryptography mechanism. This mechanism confirms the authentication of transactions carried out on the blockchain (Nakamoto, 2008; Zheng et al., 2017).

Bitcoin offered a potential solution to the dissatisfaction with the centralised financial system and the drive to find independent technological solutions. This first operational cryptocurrency started to run in January 2009 and is considered an alternative to the centrally controlled fiat currencies. It is the first decentralised cryptocurrency with a public transaction ledger that runs on a public blockchain. The idea was developed and published in a white paper in 2008 by a person, or perhaps a collective, called Satoshi Nakamoto, whose identity is still unknown. They published the paper and developed the idea behind Bitcoin. However, blockchain and the cryptography mechanism were proposed in 1976 by Whitfield Diffie and Martin Hellman. They initially proposed the asymmetric cryptography that uses public and private keys applied in the case of the blockchain (W. Diffie and M. Hellman 1976). The transactions run on a proof-of-work (PoW) hash-based chain where the information cannot be changed or erased. There are multiple benefits behind the technology, such as no double-spending, no need for third-party involvement, increased anonymity, speed and price of the transactions, and

many more (Fletcher, Larkin, and Corbet, 2021; Haber and Stornetta, 1991; Nakamoto, 2008). It is essential to mention for the dissertation and general insight into cryptocurrencies that most crypto assets run on a public transaction ledger, which means the transactions are not anonymous but pseudo-anonymous (Fletcher et al., 2021; Lu, 2018). The transaction process is better described in subchapter 2.1.2 Transactions on the Blockchain to understand the technology behind that statement.

Crypto assets have come a long way since 2009 in scale and number. According to CoinMarketCap (2022), in May 2022, there were more than 19,000 different crypto assets, 500 exchanges and a crypto market cap of more than 1.5 trillion USD.

2.1.2. Transactions on the Blockchain

Transactions on the blockchain are graphically presented in Figure 2.1, which originates from Bitcoin's white paper describing the transaction process using private and public key cryptography (Nakamoto, 2008). This type of coding is called an asymmetric cryptography mechanism and was first proposed in 1976 by W. Diffie & M. Hellman (1976).

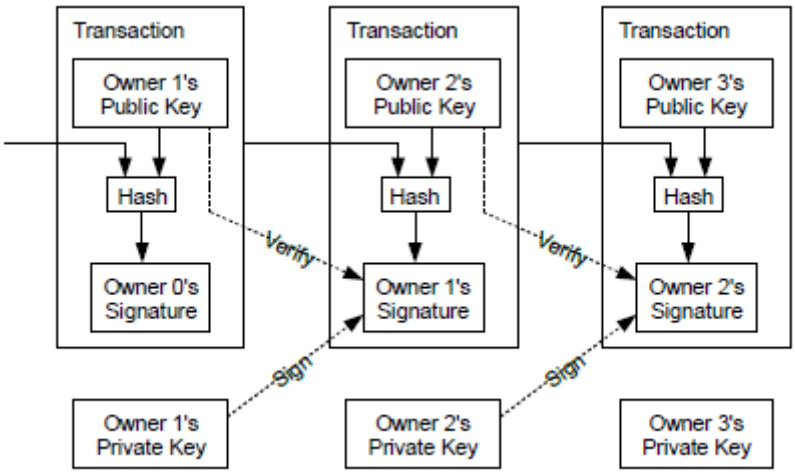

Figure 2.1 – Graphical presentation of transaction process on the Blockchain
Source: Nakamoto (2008)

The public key is visible to all on the blockchain and can be mathematically derived from the private key. However, the private key is, as the name says, private. It cannot be derived from the public key. There are four things needed for a transaction on the blockchain to be carried out (Nakamoto, 2008):

- The public key of the sender,
- The private key of the sender,
- Transaction hash, and
- The public key of the recipient.

Additionally, the blockchain needs a distributed timestamp server to avoid double spending and ensure the transactions are valid. As the blockchain is decentralised, there is no central third party to confirm the transactions, but they are confirmed by so-called consensus algorithms (Zheng et al.,

2017). There are many different algorithms, but this dissertation focuses on proof of work (PoW), the consensus used by Bitcoin and most other cryptocurrencies. The PoW cryptography method used by Bitcoin was developed by Adam Back in 1997 and is called Hashcash (Beck, 2002; Nakamoto, 2008; Zheng et al., 2017).

All the transactions are recorded on the blockchain. As long as someone has the transaction hash, they can use either public blockchain explorers or another blockchain analytical software and see all the essential information, such as (Antonopoulos, 2017):

- Time of transaction,
- Public sending address and receiving address, and
- Amount of transferred crypto assets.

Blockchain explorers are web applications that work as search engines on a specific blockchain. Not only can one check a specific transaction, but it also offers open access to addresses, individual blocks, and the transaction information for all addresses. For most crypto assets, it is freely accessible to see this information; however, making sense of it takes more effort. The explorer shows the addresses, but an observer cannot identify the account owner if the address has not been identified or recognised by service providers, such as cryptocurrency exchanges and other regulated sites (making it pseudo-anonymous). The benefit of blockchain analytical software like Chainalysis, TRM, and others is that they use various algorithms to identify addresses and offer a graphical presentation of transaction flow and exposure to different services and illicit addresses, making it easier for investigators to identify suspicious behaviour (Antonopoulos, 2017; Chainalysis, 2020a; Fletcher et al., 2021; TRM, n.d.). Considering Bitcoin was also developed to be anonymous, one issue is privacy leakage, as the transactions are very easily traceable. To improve privacy, some individuals have started using various services, one of which is mixing, which multiplies input and output transactions, and redirecting through intermediaries to cover up the flow of the cryptocurrency. This makes the transactions more anonymous, but it also makes it harder to investigate illicit behaviour (Zheng et al., 2017).

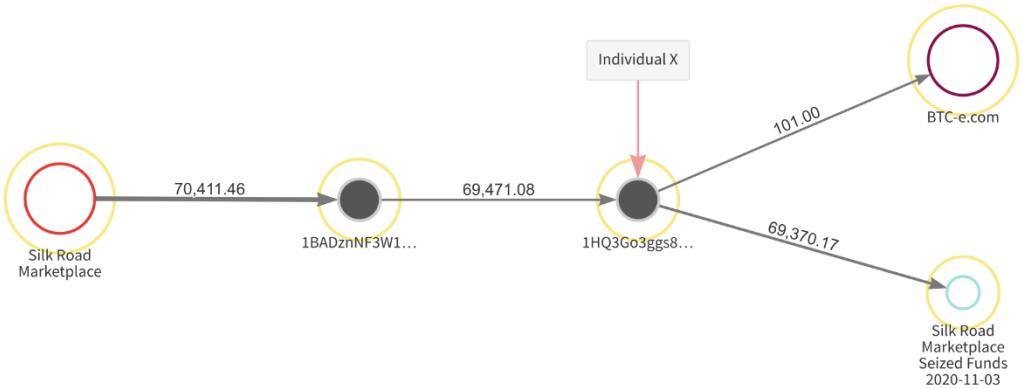

Figure 2.2 – Transaction flow – hacker Individual X who stole those funds from Silk Road
Source: Chainalysis (2020b)

Figure 2.2. presents a clear overview of the transaction flow of Individual X, a hacker who stole funds from the darknet market Silk Road and was traced by Chainalysis. The figure shows the outline in the Chainalysis Reactor software.

2.1.3. Acquiring a Crypto Asset

There are different ways of acquiring crypto assets. The initial method to obtain them was mining, which is the process of adding data blocks to the blockchain by successfully solving complex mathematical problems. Other more accessible options to acquire crypto assets are selling goods in exchange for a crypto asset, crypto ATMs and different types of exchanges. Through ATMs, one can buy Bitcoin and other crypto assets with cash. The currency is then sent to a specified wallet, meaning the individual buying the assets must have a personal crypto wallet.

When it comes to centralised cryptocurrency exchanges, users do not need a separate personal wallet, even though it is supposedly safer to withdraw money to a private wallet or cold storage. Once an individual sets up an account at an exchange, they must go through several other steps unique to the given exchange before they can start trading (Antonopoulos, 2017; Arslanian and Fischer, 2019). More information about the types of exchanges and their role is in the following subchapter, 2.2 Cryptocurrency Exchanges.

2.2. Cryptocurrency Exchanges

Cryptocurrency exchanges work similarly to other exchanges and serve as a marketplace. For example, the New York Stock Exchange is a marketplace for equities, whereas crypto assets are traded on a crypto exchange. These exchanges serve as a marketplace, but some also offer other services (Arslanian and Fischer, 2019).

There are two types of exchanges, centralised and decentralised. The latter only facilitates trading by providing an environment where buyers and sellers interact directly without an intermediary. Centralised exchanges operate, as already mentioned, similarly to stock exchanges, where buyers and sellers carry out the trades through the exchange as an intermediary and without revealing the identity of either party involved. Centralised exchanges differ in accepted currencies. There are two types (Arslanian and Fischer, 2019):

- Fiat-to-crypto – exchange users can deposit different fiat currencies (usually USD, GBP, or EUR depending on the exchange) and then convert them into crypto assets.
- Crypto-to-crypto – users can only deposit crypto assets and trade with other crypto assets. An individual has different options for acquiring crypto that can be sent to this type of exchange, which is described in subchapter 2.1.3. Acquiring a Crypto Asset.

The idea behind crypto is that it would be decentralised without centralised organisations collecting information, thus making it more anonymous and less prone to be influenced and manipulated by a third party. However, cryptocurrency exchanges reinforced their role in the market with the rise of popularity in crypto and the demand for places to exchange trade and stake assets. However, the rise of crypto asset popularity also meant the proportional rise of illicit behaviour. Whereas banks and

other traditional financial institutions have developed thorough screening processes for new clients and adhere to strict anti-money laundering (AML) regulations, that is not yet the case in the crypto industry. The latter means that the landscape of AML and compliance for cryptocurrency exchanges still needs to be more thoroughly developed. One issue limiting a uniform framework is that the regulation differs in different countries. Exchanges offering services under different jurisdictions must also follow those regulations to avoid penalties. For that reason, exchanges started to lead more comprehensive Know-Your-Customer (KYC) protocols similar to those employed by other financial institutions to comply with legal requirements. KYC protocols ensure the data is collected from the client and that risk is minimised (Arslanian and Fischer, 2019).

2.2.1. Challenges and Risks for Cryptocurrency Exchanges

Cryptocurrency exchanges are part of the financial technology (FinTech) landscape, and like many other fast-growing companies in the sector, they are facing unique challenges. A comprehensive research study recognised the challenges presented in Table 2.2. Some of the main organisational challenges at the market level include the lack of coordination and limited cooperation within the industry (Fosso Wamba et al., 2020).

Table 2.2 – Practical classification of FinTech challenges

Area	Challenge
Organisation	Lack of coordination amongst competing institutions
	No individual gains in competitive advantage
	Reluctance to agree on standards
	Weak/lack of incentives
	Network structure of banking
	Intellectual property concerns
	Limited cooperation between industry participants
Culture	Rapid transformation of financial systems
Regulation and governance	Regulation concerns about access to banking platforms
	Compliance issues
	Monitoring and enforcing increasingly demanding regulatory requirements on fast-changing, rapidly growing, and cross-border markets
Security and privacy	Susceptible to cyber attacks
Technology and standards	Robust infrastructure needed to support Fintech
	Limited system stability
	Resilience and security concerns

Source: Adapted from Fosso Wamba et al. (2020)

Some of the main risks for the FinTech businesses are related to potential financial losses due to fast-moving changes in regulation, not adhering to compliance measures, and damages done by illicit

behaviour by stakeholders or third-party individuals/organisations – cyber-attacks can be used as an example of an exchange that has historically been very private about its users and offered very lenient KYC procedures, allowing users only to provide an e-mail to set up and use their account. That changed in 2021 after Binance faced losing the authorisation to operate in a few countries due to the risk associated with the use of the platform for illicit activities. Another exchange, BTC-e, was shut down, and responsible individuals were arrested. The exchange did not carry out sufficient KYC procedures and was suspected of money laundering due to ransomware attacks. Considering the repercussions of not complying with regulations and having strict AML policies, most centralised cryptocurrency exchanges nowadays carry out more elaborate KYC processes (Arslanian and Fischer, 2019; Berwick and Wilson, 2022).

2.3. Fraud

The previous subchapters covered the basic information about blockchain and cryptocurrency, followed by this chapter that offers information on fraud and money laundering. These topics cover the basics of understanding how cryptocurrencies can be used for money laundering and how to detect it with predictive analysis, which is the topic of the dissertation.

"Fraud is both an illegal act and a criminal offence. It is intentional, deliberate, and purposeful. Fraud is underlined by deceit, concealment, violation of trust, or use of dishonest means through which possession (e.g. money, property or a legal right) is deprived of the victim. It is an offence that causes injury, material or otherwise to the victim." (Onwubiko, 2020). This basic definition by Onwubiko (2020) gives a good foundation, which will be further developed throughout this chapter.

There are many definitions of fraud, but two essential fraud-related triangles provide the basic definitions of the conditions that facilitate fraud and actions that constitute a fraud action.

Financial fraud is an illicit act that usually results in financial advantage. It is an act that has been part of processes involving finances for centuries. There have been many reports throughout history, one of which includes a financial statement fraud report as early as the 17th century. There is a great deal of historical data; whereas the types of fraud might have changed significantly with new financial instruments, technology, and the internet, there are still some commonalities. A helpful outline in terms of criteria that facilitate fraud can be seen in the fraud triangle shown in Figure *2.3*, which focuses on the perpetrator of the action and consists of (Dorminey et al., 2012):

- Financial pressure – a financial problem that results in a motive to commit fraud;
- Opportunity – a perception that there is an option to commit fraud without getting caught; and
- Rationalisation – the individual, carrying out a fraudulent act tries to rationalise the act as an exception that still falls within their moral values.

The fraud triangle was enhanced by (Wolfe and Hermanson 2004) to include the fourth criterium, Capability. The authors suggested that even if the three factors from the fraud triangle definition are covered, it is not necessarily the optimal situation for fraud action to occur. The three factors create an environment where it is possible to carry out the action. However, the individual needs to be capable of recognising the opportunity and performing the act (Dorminey et al., 2012).

Figure 2.4 presents the second fraud-related triangle, the triangle of fraud action, which outlines the actions needed for the fraud to transpire. The fraud action triangle, contrary fraud triangle, does not focus on the perpetrator but on the three actions of the crime carried out, which are (Dorminey et al., 2012):

- The Act – execution and method of the fraudulent act;
- Concealment – the actions are taken to hide the fraudulent activity; and
- Conversion – masking the origin of illicit profits to appear legitimate.

Figure 2.3 – Fraud triangle

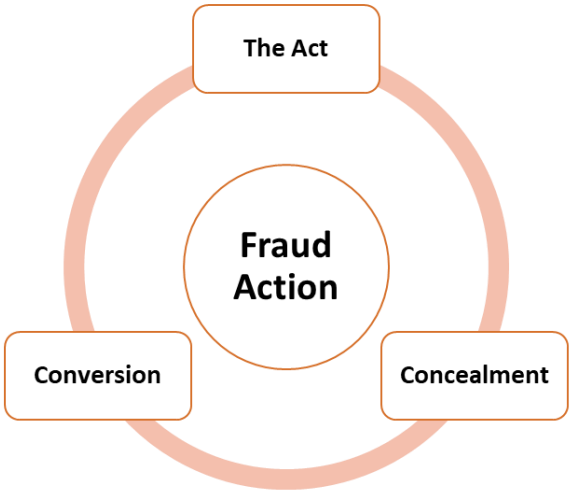

Figure 2.4 – Triangle of fraud action

Source: Figure 3 and Figure 4 adapted from Dorminey et al. (2012)

There are different types of fraud based on seven features of a framework proposed by (Onwubiko 2020). These features are presented in Figure 2.5 and include what the author considers the fundamental characteristic of every fraud. The features describe the fraud in terms of the channel, magnitude, frequency and more. These features can be integral when first trying to identify and analyse the fraud actions. It gives a broad overview of the fraud structure and elements.

Channel	Online Frauds	Occur when electronix or realtime channels such as web, mobile or telephony are used to commit fraud.
	Offline Frauds	Occur throught the use of physical channel.
Boundary	Internal Frauds	Originate inside the system boundary.
	External Frauds	Originate ooutside the system boundary but may be aided by entities inside the system boundary.
Entity	Human Frauds	Carried out or executed by a human.
	Automated Frauds	Executed by automated systems, e.g. endpoints, Bots, scripts, RPA).
Intent	Deliberate Malicious Frauds	Results from intentional decision carried out by an entity (human or Bot) with malicious objective usually by threat actors e.g. organised criminals, malware writers.
	Deliberate Non-Malicious Frauds	Result from intentional decision carried out without malicious objectives, usually by threat actors e.g. political activists, environmental campaigners, ethical researchers.
	Accidenetal Frauds	Introduced without awareness, e.g. human error, omission or negligence.
Motivation	Financial Frauds	Result in financial gains.
	Non-Financial Frauds	Result in immediate non-financial rewards.
Capability	Significant Frauds	Committed by extremely skillful and formidable threat actors who are capable of revise engineering codes and create custom codes themselves, e.g. nation-sponsored actors.
	Competent Frauds	Committed by a moderately skillful and competent threat actors, e.g. investigative journalists.
	Minimal Frauds	Commited by threat actors with basic to modest/limited capability, e.g. learners.
Persistence	Permanent Frauds	Presence is assumed to be continuous in time, e.g. account takeover.
	Transient Frauds	Presence is bounded in time.

Figure 2.5 - Elementary fraud features
Source: Adapted from Onwubiko (2020)

2.4. Money Laundering

Money laundering serves as the conversion component in the Triangle of Fraud Action and is the process of masking the money acquired with criminal activities appear as money acquired from a legitimate source (Dorminey et al., 2012). These criminal activities could include terrorist funding, child abuse, drug trafficking, human trafficking, and other illicit activities. For this to be achieved, criminal organisations use different approaches (Salehi, Ghazanfari, and Fathian, 2017). The yearly amount of money laundered globally is around 2-5% of global GDP, which is between 800 billion and 2 trillion USD. Due to the nature of concealing the origin of money, those numbers are only an estimation, as it is challenging to track and uncover all money laundering activities (United Nations, n.d.).

Money laundering typically follows three steps. The process starts with placement, continues with layering, and is completed with successful integration. A graphical outline of the stages is presented in Figure 2.6. The first stage, placement, relates to the introduction of money gained from criminal activities into the legitimate financial system. Layering, the second stage, consists of concealing the source of the money through various techniques, one of those being money mules. In integration, the

last stage, the money returns to the initial owner where the source of the funds has been hidden and can be considered legitimately acquired and freely used by the owner (Alsuwailem and Saudagar, 2020). More about how money mules are recruited and how the process is carried out is presented in the following subchapter, 2.4.1 Money Mules.

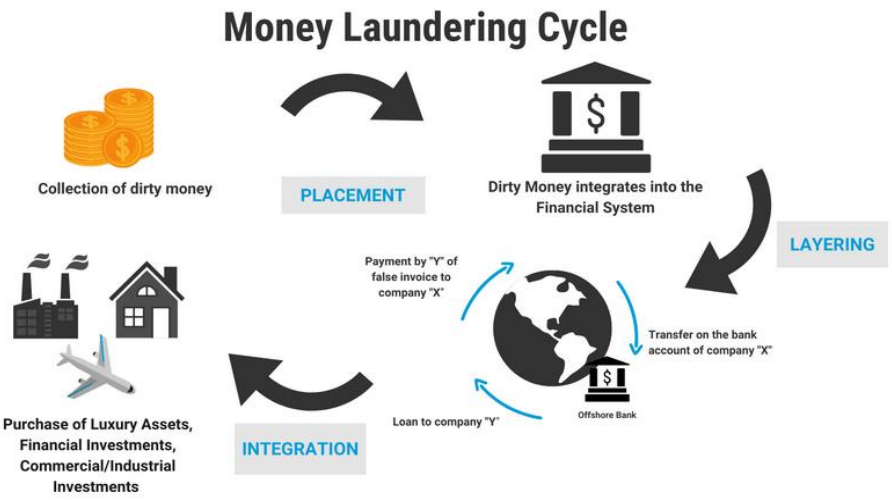

Figure 2.6 – Money laundering cycle
Source: United Nations (n.d.)

2.4.1. Money Mules

One of the approaches to legitimise money is using money mules that serve as part of the placement and layering steps in money laundering (Esoimeme, 2021). Typical money mule fraud is when an individual is encouraged to complete a money transfer for a percentage of the transfer as the payment. Individuals are often recruited where the fact of their involvement in fraudulent behaviour is omitted. They become intermediaries between “dirty money” and the guilty party laundering their money. Often, these activities are advertised and promoted on social media as a quick way to earn money. The advertising includes various approaches; some also include what seem to be legitimate job advertisements that include different forms of funds transfer. Money mules not only partake in illicit activity, but they can also face severe repercussions if caught. An individual can be charged with offences, such as concealment of criminal property, transferring criminal property, and bank fraud, to name a few. According to European Union Agency for Law Enforcement Cooperation (EUROPOL), targeted individuals for money mule activities are often those with worse economic opportunities at that moment, such as students, unemployed people, and those who recently immigrated to a new country and are looking for new opportunities. Most often, exploited people, therefore, are the ones who are the most vulnerable (Esoimeme, 2021; EUROPOL, 2021).

There are three types of money mules; according to the Federal Bureau of Investigation (FBI), these are (Federal Bureau of Investigation, n.d.):

- Unwitting or unknowing – individuals do not realise they are taking part in any illicit activity due to believing that the job, for example, is a legitimate opportunity.

- Witting – individuals who ignore red flags and partake in the activity. They might at first be unwitting but are persuaded by the financial gains.
- Complicit – willingly and actively participate in money laundering activities, earn money, and potentially recruit new money mules.

Money laundering practices in cryptocurrency and, more specifically, the role of money mules is presented in subchapter 2.5.2 Money Mules and Cryptocurrency.

2.5. Illicit Activities and the Cryptocurrency Industry

The trends in cryptocurrency crime for the year 2021 show that there has been a historic high in the amount of cryptocurrency-related crime activities; the movement in the past five years is presented in Figure 2.7. In 2021 the amount of 14 billion USD was received by illicit crypto addresses, which was a substantial rise compared to 2020, when the amount was 7.8 billion USD. However, considering the rise in the use of cryptocurrencies, the subsequent use for criminal activities is expected. The transaction volume in 2021 grew to 15.8 trillion USD, a 567% rise from the transaction volume in 2020. The rise in volume resulted in a historical low in illicit use, with only a 0.15% share compared to non-illicit transactions (Chainalysis, 2022; Europol, 2021). The total volume and share of different illicit behaviours are presented in Figure 2.7.

As presented in Figure 2.7, the main categories of illicit behaviour in the cryptocurrency industry are scams and stolen funds for 2021, followed by darknet market activity (Chainalysis, 2022).

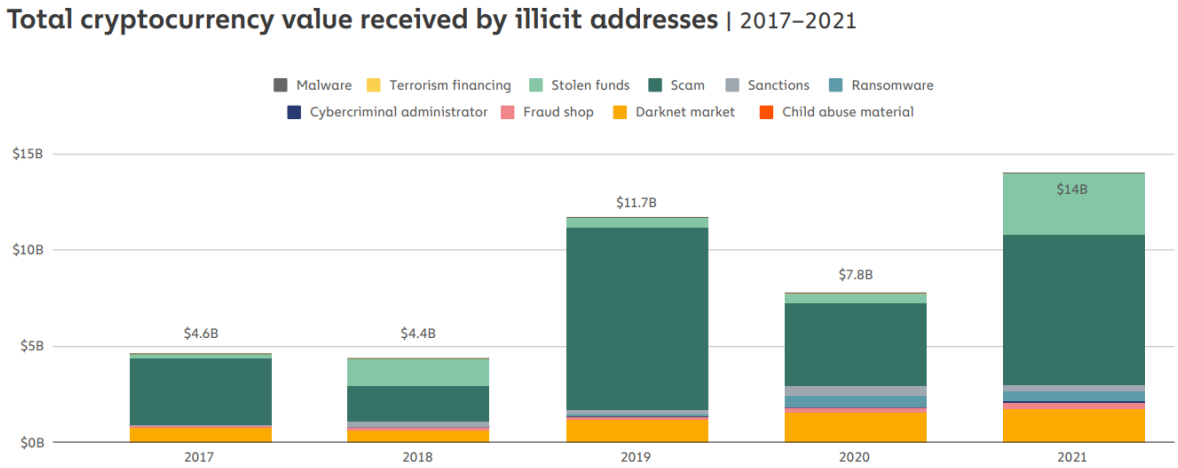

Figure 2.7 – Total cryptocurrency value received by illicit addresses
Source: Chainalysis (2022)

2.5.1. Money Laundering and Cryptocurrency

This subchapter builds up from subchapter 2.4. Money Laundering to go more in-depth on cryptocurrency-related money laundering. Cryptocurrencies are a relatively new technology with the potential for versatile implementation and use. That said, it means that there is a high probability of the new technology being exploited by different individuals and organisations (Campbell-Verduyn, 2018).

Table 2.3 – Money laundering risks posed by cryptocurrencies

General risk factors	Money laundering stages		
	Placement	Layering	Integration
Quasi-anonymity	Cryptocurrencies can be used by criminals and associations	Suspicious names, mainly if money mules are involved, that cannot be flagged	Allowing cashing out of proceeds of crime to be passed on anonymously to individuals that cannot be traced
Real-time transactions	Proceeds of crime can be transferred to another cryptocurrency to another country	Transactions occur in real-time, allowing little time to stop them if suspected of money laundering	Proceeds of crime can be moved rapidly through the global financial system and withdrawn in another country

Source: Adapted from Campbell-Verduyn (2018)

In terms of how cryptocurrencies can be used for money laundering, Table 2.3 describes general risk factors, which are quasi-anonymity and real-time transactions, and their exploitation in the three money laundering phases: placement, layering, and integration. As the transactions can be performed in real time and offer some anonymity, it substantially increases the risk of criminal activity. Not only are the transactions carried out faster than traditional financial instruments, but pseudo-anonymity also offers the perpetrators of the activity to hide ownership, resulting in the impeded tracing of criminal activity (Campbell-Verduyn, 2018).

The type of money laundering strategy sufficiently differs depending on the criminal activity. Figure 2.8 is a stacked bar graph that presents where the funds are sent depending on the activity identified by Chainalysis for 2021. The majority of scammers, for example, send the acquired funds to centralised exchanges (grey colour of the bar graph), whereas the majority of the stolen funds end up in decentralised finance (DeFi) protocols. Initially, DeFi protocols were not considered regulated as centralised exchanges, as they operate as non-custodial cryptocurrency platforms and do not hold funds but only act as an intermediary between users. However, the latest FATF guidelines from 2021 propose that the DeFi protocols also need to comply with the regulations, resulting in better KYC procedures and easier tracking of suspicious transactions and criminal activity (Chainalysis, 2021; FATF, 2021).

Cryptocurrencies are pseudo-anonymous (described in 2.1.2 Transactions on the Blockchain), which provides users with some part of anonymity. However, cryptocurrency exchanges, especially the centralised ones, follow stricter regulations and compliance, requiring more information from the exchange users, which they gain through a Know Your Customer (KYC) procedure. Exchanges that do not follow strict compliance are more prone to exploitation for illicit activities, as the users get more anonymity. As these exchanges present a higher risk and have low KYC requirements, they are regarded as riskier. For example, Chainalysis, one of the most widely used blockchain analytical software, classifies them as “high-risk exchanges.” Whenever transactions are made to or from these addresses, an alert is triggered that needs to be investigated (Chainalysis, 2022).

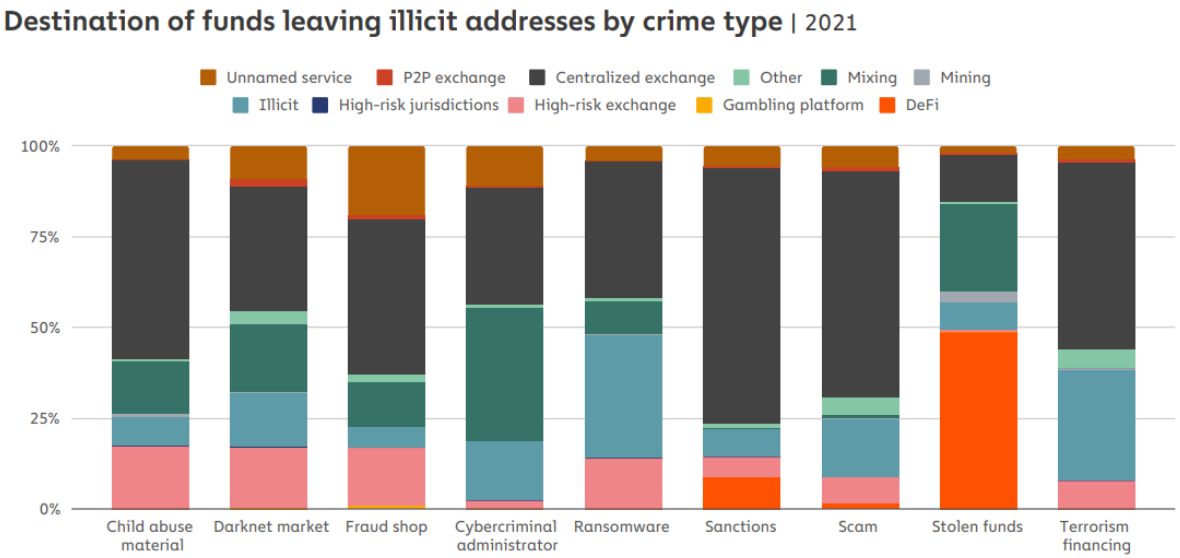

Figure 2.8 – Destination of funds leaving illicit addresses by crime type

Source: Chainalysis (2022)

2.5.2. Money Mules and Cryptocurrency

Similar to other industries, money mules are also a tool in the cryptocurrency industry that criminals use to disguise their funds or products. Cryptocurrency is different in that there are also crypto assets involved in money laundering, perhaps through the whole process or as the perpetrator of a criminal act gained crypto assets as payment or reward. Another option is crypto assets only playing a role in the second stage when the money mule exchanges the fiat money for an asset and then transfers it to an account specified by the money launderer. There are many strategies, but the process is a bit more challenging to identify because cryptocurrencies offer more anonymity than traditional financial instruments and the real-time transactions outlined in Table 2.3. Thus, when investigating suspicious activity, the placement and layering stages money mules are involved in can be carried out faster. Detection could be lacking due to the specificity of the process, which could initially be more covert with transactions of small amount. This is a common occurrence on an exchange with a large volume of transactions, making the fraud detection process more complex (Campbell-Verduyn, 2018; Fletcher et al., 2021).

2.6. Predictive Analytics and Fraud Detection

Fraud and money laundering are an ever-present threat to financial institutions that can be mitigated with different approaches. One of these approaches is predictive analytics, which is part of data science that can be applied to many disciplines where there is available data in which patterns can be uncovered. Prediction could, in general terms, mean the prediction of future events; however, in data science, it relates to the prediction of an unknown attribute built on historical or present data. There are two types of predictive analytics depending on the type of the target value. When the target value is continuous, capturing values in a limited or unlimited interval, the predictive modelling used is regression. One of the potential applications of regression in fraud detection would be predicting the fraud amount. The second type of predictive analytics is called classification and is used when the

target value is categorical. The target value can be binary (in our case – mule, not mule) or multiclass, where the target values capture multiple values (Baesens, Vlasselaer, and Verbeke, 2015; Gandomi and Haider, 2015; Provost and Fawcett, 2013). The process through which predictive models are trained is called machine learning. The machine learning process goes through different potential prediction models to identify the one that fits the data best and detects the relationships (Kelleher, Mac Namee, and D'Arcy 2020). Gandomi and Haider (2015)suggest that most methods used to find patterns were developed and meant to be used on small datasets. The constant increase in data availability and more efficient data processing facilitates the development of new methodologies and more efficient usage of old methods. More data and new methodologies offer new and potentially more precise outcomes.

Whereas existing rule-based systems for fraud detection within the financial industry might detect individual suspicious transactions and behaviour changes, they cannot adapt to more complex patterns. Machine learning becomes valuable because it can be applied to various situations and continuously adapt to new data. It enables more complex pattern recognition that can evolve through time and continue to identify suspicious activity (Zhang and Trubey, 2019). One thing is certain, according to Baesens, Vlasselaer, and Verbeke (2015), and that is that crime is an ever-evolving activity that always finds a new way to exploit the system. While fraud is evolving, so are the methods to detect it. Fraud could be detected with many different methods, including descriptive, predictive, or social network analytics. The following subchapters describe several approaches to machine learning algorithms that can be applied in the cryptocurrency ecosystem, focusing on how supervised machine learning and predictive analytics can be used for fraud detection.

2.6.1. Supervised and Unsupervised Machine Learning

Machine learning algorithms can be applied to various data-related problems, such as finding patterns in data. There are three types of machine learning: supervised, semi-supervised, and unsupervised. In reviewing the research and solutions, it is clear that the use of machine learning for AML solutions is generally based on supervised or unsupervised machine learning, depending on the type of potential fraud and the availability of data (Chen et al., 2018).

Unsupervised machine learning uses algorithms to divide data into different groups, or clusters. Clustering distributes data into groups with certain similarities depending on what we are looking for. An example of the use of clustering in the cryptocurrency industry would be blockchain analytics tools such as Chainalysis. One of the ways the analytics software is applied is by using algorithms to cluster addresses of different services. The effectiveness can be presented on a darknet market, the Silk Road, which shut down in 2013 after the owner was arrested. The United States Department of Justice managed to identify the cryptocurrency wallets connected to the "first major digital darknet market" using Chainalysis analytical software. That allowed them to identify the addresses of the wallets and seize assets worth more than 1 billion USD at the time of the seizure. The software clusters different addresses and maps out transactions that can be graphically investigated, as the blockchain is transparent and transactions can be easily tracked. The problem is identifying who owns the address; for example, if it is not connected to a centralised exchange with thorough KYC procedures, clustering can provide better insight into risky exposure (Chainalysis, 2020a; Chen et al., 2018).

Supervised machine learning works differently from unsupervised. The data for supervised machine learning includes labelled examples or training sets on which the model builds the prediction. As in the case of this dissertation, the labelled examples can be used to train the algorithm to classify all data based on the features of the training set (Chen et al., 2018; Ngai et al., 2011).

2.6.2. Supervised Machine Learning Algorithms for Fraud Detection

This chapter will focus on financial fraud detection related to cases where the data structure is comparable to the money mule fraud on the cryptocurrency exchange.

Considering that the data consists of users who did not carry out fraudulent actions and confirmed fraudulent users, the most fitting approach is to use supervised machine learning to conduct the analyses. Research shows examples of financial fraud detection, which could be done with predictive analytics using classifications and regression. Data analytics analyses large amounts of data to gain insight, whereas classification falls into the spectre of supervised machine learning (West and Bhattacharya, 2016).

The crypto industry is a relatively new sector compared to other financial instruments and institutions. For example, the banking system has been around far longer, resulting in more data and research about various fraud types and techniques to detect it. The research for this dissertation focuses on fraud at a financial institution, and for the methods used, it is helpful to investigate the approaches currently being used. Research into the most used widely methods for fraud detection demonstrated that most cases are related to bank fraud. The data structure is similar to the cryptocurrency exchange. For example, credit card frauds and bank frauds are similar in data availability. The bank, or other financial institution, has confirmed cases and can make data analyses to help detect other potential victims. The money mule or victimised party could have made a voluntary action involving them in the crime, which differs from credit card victims where the data was often stolen. However, the idea is similar in confirmed cases, and large data and information about the individuals involved enable us to use supervised machine learning techniques. Banks, other financial institutions, and some cryptocurrency exchanges have in-depth KYC procedures. They gain user data, which helps evaluate risks, prevent money laundering, protect customers, and perform various analyses. Collecting data from different sources enables a comprehensive overview of client behaviour that can be analysed for various business process optimisations (Ngai et al., 2011).

The fraud occurring in the cryptocurrency industry is a rather recent development. Using existing approaches to fraud in other financial institutions can be useful for solving similar issues. For example, West and Bhattacharya (2016) and Ngai et al. (2011) list several types of financial fraud. Similar to the money mule issue in the dissertation, the cases of bank fraud, more specifically credit card fraud, had some common points. The similarities are in terms of available data and problem structure. Both articles list some of the most often used techniques for classification prediction. The most widely used methods for fraud detection in various financial institutions included logistic regression, decision trees, and various ensemble methods like random forest and boosting, such as XGBoost, which use multiple models to carry out estimations (Baesens et al., 2015; West and Bhattacharya, 2016). Other methods also included the Support vector machine and neural network. The algorithms used for the dissertation are better described in chapter 3.3 Modelling Algorithms.

2.6.3. Common Machine Learning Issues in Fraud Detection

There are several issues when trying to obtain a prediction model. The most common issues researchers face and how to minimise the effect on the model are described in the following subchapters. These issues include data imbalance, false positive and false negative classifications, and overfitting and underfitting.

2.6.3.1. Data Imbalance

Detecting fraud instances can often present as a rare event, meaning that in large data, there could potentially be only a few fraud instances representing the minority class. In contrast, the majority class would be instances that are not fraudulent. However, identifying a fraud occurrence is vital for financial institutions, meaning that correct classification results are integral for optimal operations. Several methods help improve classification results when dealing with imbalanced data. When doing the pre-processing of the data, two steps within some operations can be performed that would improve the results: resampling of the data and feature selection. Feature selection aims to select a subset of variables from the dataset that would ensure optimal model performance and move irrelevant variables, thus reducing the risk of the minority class being overlooked. Resampling can be done with both over-sampling, under-sampling, and a combination of both. Over-sampling is when the minority class instances are replicated to create more samples synthetically. One of the most widely used techniques for oversampling is SMOTE (Synthetic Minority Oversampling Technique). Under-sampling removes samples from the majority class to minimise the class imbalance. Random Under-sampling is one of the most widely used techniques for under-sampling (Guo et al., 2017).

Another integral factor to consider is the selection of the most appropriate methods for the algorithm. If the chosen algorithm is suitable for balanced data, applying it to imbalanced data results in underwhelming results (Guo et al., 2017).

2.6.3.2. False Positives and False Negatives

False positive and false negative cases can cause considerable constraints on a business depending on what we are trying to predict. Especially sensitive are predictions that affect people, for example, in the health sector. Models predicting different diseases need to be very accurate, and the misclassification error needs to be low to avoid endangering people's lives. For fraud detection, having a good prediction model could fundamentally improve the efficiency of fraud prevention. However, having many false positive cases causes a constraint, as it would mean many resources are needed to investigate the predicted positive fraudulent events (Baesens et al., 2015; Kelleher et al., 2020).

Additionally, customers would not appreciate being marked as fraudulent after having their accounts blocked due to false classification. Customer dissatisfaction could cause higher customer churn. Implementing a model means that follow-through is essential to ensure the company finds many positive fraud cases. If there are many false negative cases, it would mean that only implementing that specific model exposes the company to various risks from fines to a poor reputation. It is integral to develop a well-performing model with a low number of false positive and negative predictions, or perhaps not implement the model into the processes if the performance is not good enough to avoid

these negative consequences. The model is good enough based on the individual case and business type and cannot be predefined (Baesens et al., 2015; Kelleher et al., 2020).

2.6.3.3. Overfitting and Underfitting

Two issues that are often encountered in data modelling are overfitting and underfitting. Suppose the modelling is performed on split data. In that case, overfitting can occur when the model is too complex and memorises the training data instead of identifying meaningful relationships. This results in poor generalisation of new data and making good predictions. The model fits all data points but is too complex and can be sensitive to data noise. Underfitting occurs when the model is too simple and does not explain the variance (Baesens et al., 2015). A simple outline of the model fit is depicted in Figure 2.9, where the model shows the relationship between income and age. Underfitting is shown in graph (b), and overfitting is shown in graph (c) within Figure 2.9.

Different measures can be applied to avoid both issues when setting up the algorithms. The most basic is to split the dataset into training and test sets. Regarding rare events and modelling, the ROC can be somewhat sensitive, resulting in significant variations in the AUC. Splitting the data ensures we gain a more complete picture of the performance. The training set is used to train the machine learning algorithm, whereas the test set is used to test the prediction model (Kelleher et al., 2020; Zhang and Trubey, 2019).

Additionally, the cross-validations of the training set can be applied where the data is split into K folds. The model is trained on the K-1 folds, and model performance is estimated with the test fold. When splitting the data, the stratified split can be applied, which ensures the same proportion of majority and minority classes in all folds (Baesens et al., 2015).

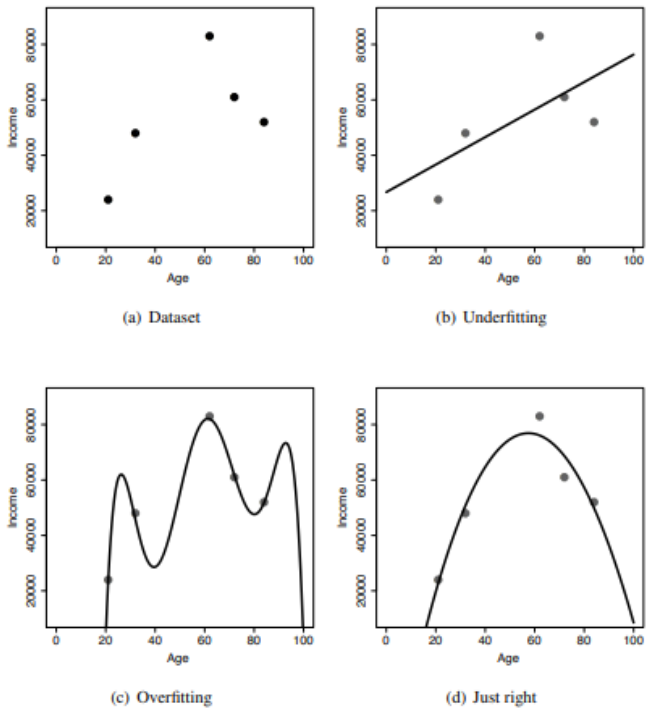

Figure 2.9 – Potential over and underfitting of a model predicting income
Source: Kelleher et al. (2020)

2.6.4. Metrics for Classification

Many different metrics can be used when evaluating predictive analytics model performance. This subchapter focuses on the two main methods used to evaluate the money mule models: confusion matrix and a Receiver Operating Characteristic (ROC) Curve. Additionally, the individual parameters to calculate ROC and the confusion matrix were used.

2.6.4.1. Confusion Matrix

One metric known to be very useful and widely used when classifying events is the confusion matrix, the outline of which is in Figure 2.10. In classification, the predicted event is either correct or not, so it is either negative or positive. The matrix encodes the results of the tested prediction model; they are tested where 0 means negative and 1 means positive. There are four elements related to the classification of the actual target value and the predicted classification. These four elements are (Kelleher et al., 2020):

- True positive (TP) – the events the model predicted correctly, meaning the model predicted a positive classification and the target value was indeed positive.
- True negative (TN) – the events the model correctly predicted as negative.
- False positive (FP) – the events the model predicted to be positive, but the target value was negative.
- False negative (FN) – the events the model predicted to be negative, but the target value was positive.

		Prediction	
		0	1
Actual value	0	TN	FP
	1	FN	TP

Figure 2.10 – Confusion Matrix
Source: Adapted from Kelleher et al. (2020)

The prediction model is good if the values on the diagonal where there are true positives and true negatives are high, meaning that the number of correctly predicted events is high. The other diagonal with false negatives and false positives shows the mistakes the model made when predicting the target values. From the confusion matrix, it is easy to calculate the misclassification rate and classification accuracy (Kelleher et al., 2020):

$$misclassification\ rate\ (MR) = \frac{(FP + FN)}{(TP + TN + FP + FN)} \quad (1)$$

$$classification\ accuracy\ (CA) = 1 - MR = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (2)$$

Source: Adapted from Kelleher et al. (2020)

Misclassification or an error rate is the proportion of the events that were predicted incorrectly, including false positives and false negatives, compared to all the predictions. The opposite is classification accuracy, calculated as the proportion of events predicted correctly compared to all the predictions. The values can sit between 0 and 1; for accuracy, the higher the score, the better the result, whereas for a misclassification rate, the opposite is true, so a lower value is better (Kelleher et al., 2020).

Several other performance parameters can be calculated based on values in the confusion matrix. These are the true positive rate (TPR), the true negative rate (TNR), the false positive rate (FPR), and the false negative rate (FNR). The range of these parameters is between 0 and 1. For the true positive and true negative rate, the higher the value, the better the model. The opposite is true for false positive and false negative rates. TPR is also known as recall or sensitivity and is the probability that the model predicts a positive outcome for observation when the outcome is positive, while TNR is known as specificity, which is the probability that the model predicts a negative outcome for observation when the outcome is negative. One way to visualise these two metrics is by creating a ROC curve, which stands for the "receiver operating characteristic" curve, specifics of which are presented in the following chapter (Kelleher et al., 2020).

$$TPR, sensitivity, recall = \frac{TP}{(TP + FN)} \tag{3}$$

$$TNR, specificity = \frac{TN}{(FP + TN)} \tag{4}$$

$$FPR = 1 - specificity = \frac{FP}{(TN + FP)} \tag{5}$$

$$FNR = \frac{FN}{(TP + FN)} \tag{6}$$

Source: Adapted from Kelleher et al. (2020)

2.6.4.2. Receiver Operating Characteristic (ROC) Curve

A literature review showed that the most used metric from imbalanced data was the receiver operating characteristic curve (ROC curve). For model performance comparison, Area Under the ROC Curve is often used. ROC is a performance measure that uses prediction scores for calculations. An ROC curve has FPR on the x-axis and TPR or specificity on the y-axis. The curve is highest when it reaches the top left corner, where AUC is 1, which means that the TPR is 1 and FPR is 0. In this case, the model predicts all the events correctly. At an AUC of 0.5, the model does not distinguish between a positive and negative case, meaning TPR and FPR are equal. The higher the measure's value, the better the model's performance. Models with values below 0.6 are generally considered weak, whereas those above 0.7 are considered strong (Kelleher et al., 2020). The outline of an ROC curve is shown in Figure 2.11, which includes two curves, one with an AUC over 0.9 and one with an AUC below

0.7. An AUC of 0.5 is shown as a diagonal, dashed line. In the figure, the model with the blue line, with an AUC of 0.9141, performs better than the one with the orange line, with an AUC of 0.6998.

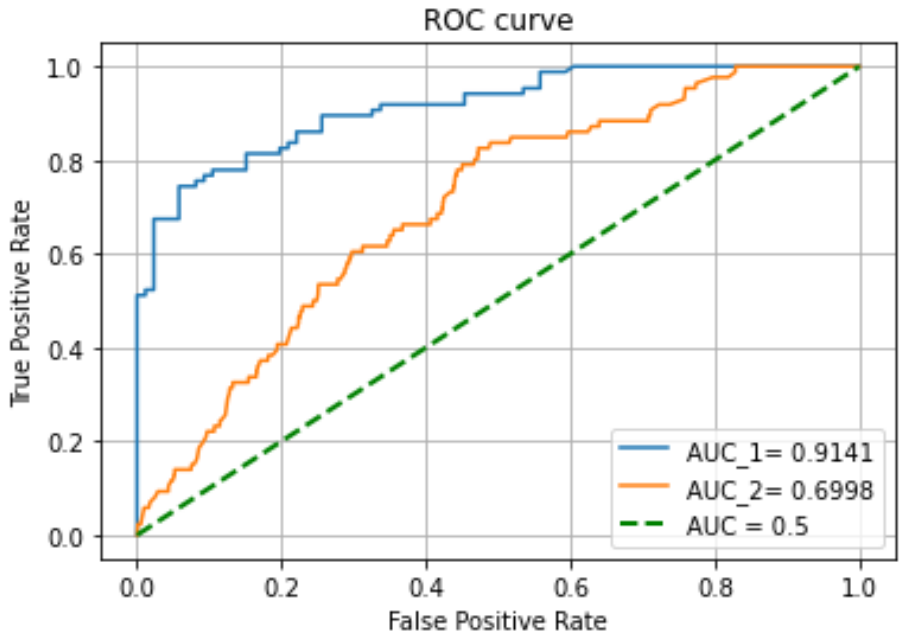

Figure 2.11 – ROC curve example

Source: Own work

A review of existing research by West and Bhattacharya (2016) showed that the results for the prediction models in fraud detection for bank fraud reach high results. Accuracy for some credit card cases was above 95%, whereas sensitivity was between 42% and 82% for the random forest. Baesens, Vlasselaer, and Verbeke (2015), collected performance benchmarks list where they list research that has achieved from 90% to 98.6% AUC values for credit card fraud analyses. What is common for these research papers is that the amount of data analysed was in tens of thousands or even millions of events in the dataset. However, the proportion of positive cases in those cases was between 0.1% and 1%.

3. METHODOLOGY

The main goal of the research was to develop a prediction model that would be applied to a real scenario: screening on the cryptocurrency exchange to detect potential money mule cases earlier to minimise the damages these fraudulent cases can cause. The dataset used originated from a selected cryptocurrency exchange company.

The data analysis part of the research followed the CRISP-DM methodology described in subchapter 3.2 CRISP-DM Methodology. The approach follows the model development from the first stage, Business Understanding, to the last stage, Deployment. The methodology fits the situation, as the dissertation aims to solve a real-life issue. A wholesome approach to developing a solution enables a better understanding of the overall situation.

3.1. Selection of the Platform/Environment

The data analysis was carried out using Python with Spyder as an integrated development environment (IDE). Python is a valuable tool for analyses and has the potential to be applied to many situations. Therefore, using it for fraud detection seemed like the ideal fit. Python is also one of the most common languages used for predictive analytics (Kelleher et al., 2020). Choosing Spyder as the IDE was based on personal preference and convenience due to the software being free access. Some of the visual presentations of the dataset and results were prepared with the data visualisation tool Tableau.

3.2. CRISP-DM Methodology

The methodology used as an approach to the research is CRISP-DM. It is a data mining methodology that follows a six-step process (Chapman et al., 2000; Kelleher et al., 2020). The model was developed for data mining projects. However, as the field has many similarities with predictive analytics and offers a structured approach, it can also be applied to other projects (Kelleher et al., 2020).

The methodology consists of six stages, the flow of which is presented in Figure 3.1. The figure shows three dotted lines that indicate that the stages are interlinked. These stages are (Chapman et al., 2000; Kelleher et al., 2020):

- Stage 1 – Business Understanding
  The first stage focuses on understanding what the research intends to achieve and the project's requirements. Understanding what business issue the predictive analytics model aims to solve is essential.
- Stage 2 – Data Understanding
  The data is collected in the second stage, Data Understanding, and essential information and issues are presented. This stage gives a basic understanding of the data and information on what needs to be processed in the third stage, Data Preparation.
- Stage 3 – Data Preparation
  In the third stage of the approach, the data is cleaned and transformed to gain a dataset with no issues affecting Modelling, which is carried out in the fourth stage.
- Stage 4 – Modelling

Modelling techniques are optimised to reach optimal performance. It is the phase where different machine learning algorithms are used to build predictive models.

- Stage 5 – Evaluation
  The Evaluation stage ensures that the model was developed according to the objectives set in the Business Understanding stage and that the most appropriate model is chosen.
- Stage 6 – Deployment
  The last stage of the CRISP-DM approach is Deployment, which can be carried out in different ways. This stage usually refers to implementing the developed models into the company's processes. Stage 6 is, in a way, the last stage, but in practice, the information gained can facilitate or present a need for new processes.

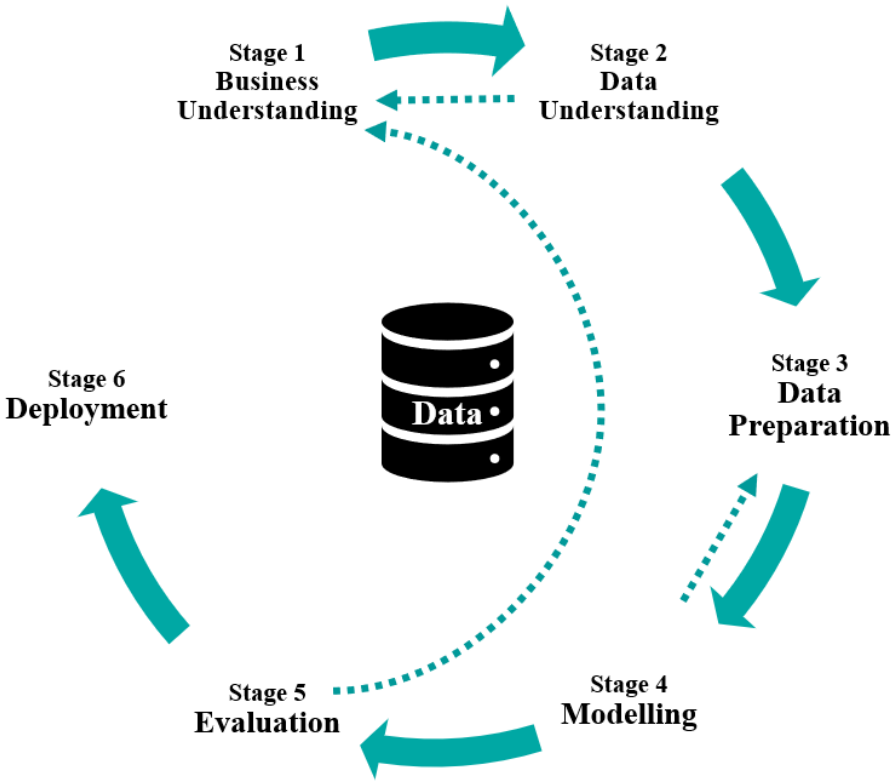

Figure 3.1 – CRISP-DM methodology sketch
Source: Adapted from Chapman et al. (2000)

The structure of the CRISP-DM methodology enabled an efficient fraud examination process by providing an understanding of the fraud process and its effects on the business, solution development, and optimisation of the model (Kelleher et al., 2020).

3.3. Modelling Algorithms

Several machine learning techniques were used to find the optimal predictive model for this research. These algorithms include logistic regression and decision tree, followed by an ensemble method called random forest and boosting algorithms XGBoost and LightGBM.

3.3.1. Logistic Regression

Logistic regression has been used frequently in various fraud detection cases and has shown good prediction performance. The model performs poorly if the data is imbalanced and there is a high correlation between predictor variables (multicollinearity). Therefore, for the model to perform better, it is essential to remove highly correlated variables and preferably have balanced data (Baesens et al., 2015).

Outputs of the logistic regression model are always between 0 and 1 and offer the probabilities of the target variable classification. Equations (7) and (8) show how the probability is calculated depending on the value of Y (target variable), which can be 1 (true) or 0 (false), as it is presented in the case of classification. $X_1, \dots, X_n$ are the predictive variables assigned to each predictive variable as a coefficient. *e* is Euler's constant, which is the base for natural logarithms and is rounded up to around 2.7183. Equation (9) is the logistic regression function defined with log odds, also known as logit transformations (Baesens et al., 2015; Hastie, Tibshirani, and Friedman, 2009).

$$P(Y = 1|X_1, \dots, X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \tag{7}$$

$$P(Y = 0|X_1, \dots, X_n) = 1 - P(Y = 1|X_1, \dots, X_n) = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \tag{8}$$

$$log\left(\frac{P(Y = 1|X_1, \dots, X_n)}{P(Y = 0|X_1, \dots, X_n)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \tag{9}$$

$$\prod_{i=1}^{n} P(Y = 1|X_1, \dots, X_n)^{Y_i} P(Y = 0|X_1, \dots, X_n)^{1-Y_i} \tag{10}$$

Source: Adapted from Hastie et al. (2009)

The β parameters are estimated with the formula in Equation (10), which estimates maximum likelihood. The parameters chosen take values that maximise the probability of attaining the target value (Baesens et al., 2015; Hastie et al., 2009).

3.3.2. Decision Tree

Decision trees are tree-like structure algorithms that detect patterns in data. Decision trees are often used due to easy interpretability and fast application in the case of issues that are not too complex. The model can include numeric as well as categorical predictor variables and is not sensitive to outliers and missing values (Hastie et al., 2009).

There are three parts of decision trees connected with branches (Kelleher et al., 2020):

- Root node (starting node),
- Interior nodes,
- Leaf nodes (termination nodes).

A sample decision tree is shown in Figure 3.2. The root or starting node is the “age” attribute, and “size_of_max_deposit” is an interior node. The bottom rectangles are the termination nodes that show the classification “yes/no.” If the instance is fraudulent, it is classified as “yes,” and “no” if the instance is not fraudulent. The tree illustration presents a very simple decision tree that predicts an instance is fraudulent in two cases:

- The user is aged 30 or less and made a maximum deposit of less or equal to 351.8 USD.
- The user is over 30 and made a maximum deposit of less or equal to 1596.6 USD.

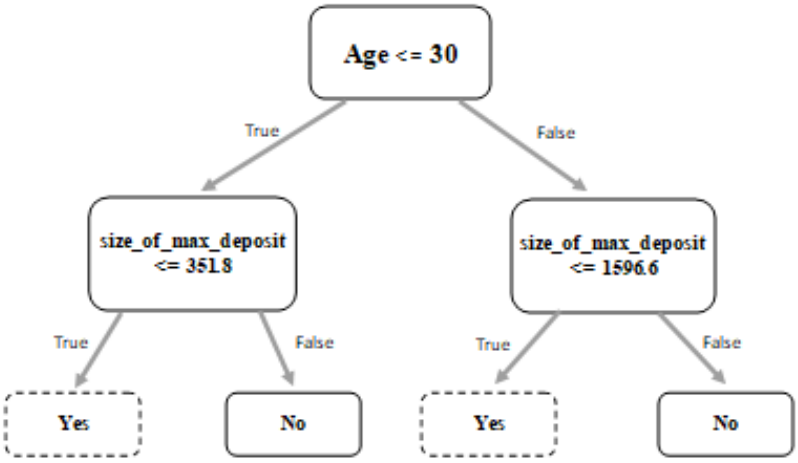

Figure 3.2 – Sample decision tree
Source: Own work

The decision trees are a high-variance tool that is very sensitive to changes in training data, which means that the outcome model changes significantly even for a small change in data. However, they offer a good base for ensemble methods, as variance and inaccuracy can be minimised by applying other methods (Hastie et al., 2009; Kelleher et al., 2020).

Machine learning algorithms that use decision trees are prone to overfitting. That could occur due to sampling variance, meaning that the target variable distribution differs between the training and the complete set. More reasons include noise in the data, as the decision trees tend to isolate noisy events and split the data based on features irrelevant to the overall situation. As the tree depth increases, there is a higher probability of overfitting, as the data is split into more subsets with each node. One of the approaches to avoid overfitting is tree pruning, which removes subtrees considered irrelevant with termination nodes that make a classification based on the feature level of the majority target within the merged subtrees being removed. An effective yet simple technique for pruning is to set an early stopping criterion when defining the model parameters. For example, the decision tree in Python has a parameter “max_depth,” the value of which will stop the tree at a set depth. While pruning affects the consistency of the model negatively when fitted to training data, it improves the model’s generalisation, making it more applicable for larger implementation (Baesens et al., 2015; Kelleher et al., 2020).

3.3.3. Model Ensembles

Model ensembles are prediction models built of different models of the same algorithm. The idea is that combining models could potentially result in improved outcomes. There are two ensemble approaches, bagging and boosting. One of the differences is that bagging can run parallelly, and the final prediction is merged, while boosting works through iteratively adding models to the ensemble. These approaches can be time-consuming and harder to interpret than decision trees; however, they decrease the variance and increase the robustness of prediction. Ensemble models also provide better results when dealing with imbalanced datasets, whereas decision trees and logistic regression work better on balanced datasets (Guo et al., 2017; Hastie et al., 2009).

Random forest is a useful ensemble method for classification or regression that builds a forest from decision trees using bagging and subspace sampling. Bagging is when a random sample of the dataset (sample is the same size as the dataset and replaced) is used to train each model that is part of the ensemble. Subspace sampling is a sampling technique where each sample consists of a subset of predictive features chosen randomly. The number of variables in the subset is for classification equal to the square root of the number of all variables. Random forests consist of de-correlated trees (enabled by subspace sampling, as the variables are chosen randomly) that are bagged and then averaged to get the final prediction (Hastie et al., 2009; Kelleher et al., 2020; Zhang et al., 2017).

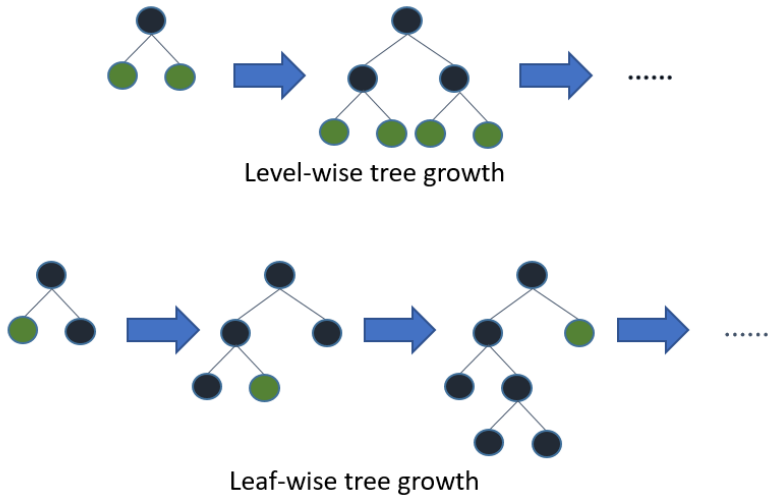

Figure 3.3 – Level and leaf-wise tree growth
Source: Microsoft (n.d.)

Boosting is an iterative process when each new model focuses on instances the previous model incorrectly classified. This is done by adapting the weights for each instance. The model starts with assigning weights to the instances as $w_i \geq 0$ is set to $\frac{1}{n}$ where w is the weight for instance i, and n is the number of all instances. The boosting process improves the model by changing the weight with each iteration, adding new models with increased weights for misclassified instances and decreased weights for correctly classified instances. The two models used in the dissertation go a step further than the basic boosting model. The eXtreme Gradient Boosting (XGBoost) algorithm and Light Gradient Boosted Model (LightGBM) are gradient boosting models that, like boosting, work in an iterative way. However, gradient boosting is more extreme, as it does not change the instances'

weights. However, each new model is set to directly correct the misclassified instances from the previous model. The model starts with a base model, and the ensemble is built with the iterative addition of new models. The process stops when it reaches the set number of models in the ensemble (Hastie et al., 2009; Kelleher et al., 2020).

LightGBM was developed by Microsoft, where the boosting of trees is carried out vertically (leaf-wise). In contrast, XGBoost trees are developed horizontally (level-wise) – the difference between the two is presented in Figure 3.3. This enables the model to perform faster, and the results are more accurate than XGBoost. LightGBM offers more options for hyperparameter optimisation, which means it can be applied to many issues. Both algorithms are prone to overfitting with increased complexity. Therefore, optimising hyperparameters is essential to ensure quality results (Microsoft, n.d.; XGBoost developers, n.d.).

4. RESULTS AND DISCUSSION

Chapter 4 Results and Discussion consists of the six stages of the CRISP-DM methodology described in 3.2 CRISP-DM Methodology. The process of the data analysis and results from the steps taken to follow the progression of the methodology are described in-depth in the following subchapters.

4.1. Stage 1 – Business Understanding

Fraud is an ever-present threat in business, but the risk is rapidly growing due to technological improvements providing new opportunities for fraudulent behaviour. Fraud can result in fines for businesses as punishment for not carrying out AML practices well enough. One of the highest paid settlements in the past decade was by HSBC bank, which paid a 1.9 billion dollar fine connected to money laundering. The amount is substantial, and it is only one of many recent settlements (Chen et al., 2018). There is also a danger of the closure of companies, such as BTC-e, a cryptocurrency exchange suspected of money laundering. A less severe constraint could be that banks might stop operations with the company due to higher risk scores linked to fraud exposure (Arslanian and Fischer, 2019). As the crypto industry is a relatively new and fast-evolving environment, it also means that the regulation is evolving as fast as the industry. Therefore, companies need to find an effective and efficient way of dealing with AML issues so that the operations do not result in financial damages to the business. Hence, implementing more comprehensive and automated fraud detections systems is key.

With this research, I set out to develop a predictive analytics model that would enable better detection of money mules on the cryptocurrency platform, thus minimising the risk to the business in terms of fines, risk scores, and reputation. Well-operating compliance practices can help with corporate liability mitigation.

Predictive models can be used in different ways by financial institutions. It could be implemented and automated if the model performance is good enough based on company-specific objectives. An automated model could help detect potentially fraudulent users in real-time. For the model to be implemented independently, the performance needs to be very good because if a poor-performing model is implemented, it could cause too many constraints on the business and the need for more resources to deal with consequences. In this case, poor-performing models include models that result in many false positive and false negative cases. Generally, it is more relevant for financial institutions and the selected company to detect all positive cases that would be achieved with no false negative cases (Arslanian and Fischer, 2019). The number of false positive cases is usually the second most relevant issue, but it can sometimes overpower the first one. That could happen if too many false positive cases result in the model causing too many constraints.

4.1.1. Problem Identification and Project Objective

Illicit behaviour of certain users that resembled money mules were identified by the selected company's fraud department as part of control operations. The source of confirmed fraud cases was a third-party stakeholder, and the type of fraud being money mule related was identified by the company's fraud department. Those users were involved in fraudulent practices. The department was

the initiator of further data analysis to find patterns that could potentially result in measures that would minimise further exploitation of the systems.

With the development of a machine learning algorithm, presented research aimed to minimise confirmed money mule cases with implemented systemic fraud detection algorithms. This would alert the system of suspicious behaviour before the layering stage of the money laundering cycle could be successfully carried out. The money could then be integrated into "clean" operations. I conducted supervised machine learning analyses to identify what would be considered suspicious.

Money laundering is implemented in different ways, and one of the risks is that even if we identify some money mule patterns, the fraud process will change by the time we accomplish that, making our detection process irrelevant. The project's objective was to identify patterns based on the different measures that could be applied. These measures would include alerts that would be triggered based on user behaviour. For example, users with identifiable features could be stopped before they made withdrawals through the exchange. This would prevent the third stage of money laundering, the integration of money, from being carried out. This would enable the detection of fraudulent users that would subsequently be reported to the institution of the relevant jurisdiction. For example, in the United States, that would be Financial Crimes Enforcement Network (FinCEN) (Arslanian and Fischer, 2019).

4.1.2. Money Mule Problem Description

The cryptocurrency exchange where the data originated is a fiat-to-crypto exchange, where the money mules were depositing fiat funds to their accounts, offering additional information for data analysis to find relevant connections.

Using the table presented in subchapter 2.3 Fraud**,** the features of the money mule case used explicitly in dissertation research can be laid out. Figure 4.1 shows an elementary fraud features outline adapted to the researched case of money mules on the cryptocurrency exchange. The illicit behaviour in question was carried out by humans (entity), the location was online (channel), and the individuals involved were external (boundary), as the internal involvement from the exchange was not present. Concerning the intent, on one side, we could argue it was deliberate malicious fraud as the main, the most probable criminal party was using money mules to launder money. However, regarding the individual money mules involved, they could have been victims of scams or perhaps did not know what they were involved in, so it could potentially be accidental fraud. The motivation for the fraud is financial gain, as the money mules get compensation, and the party laundering money gets financial gains if the process is successful.

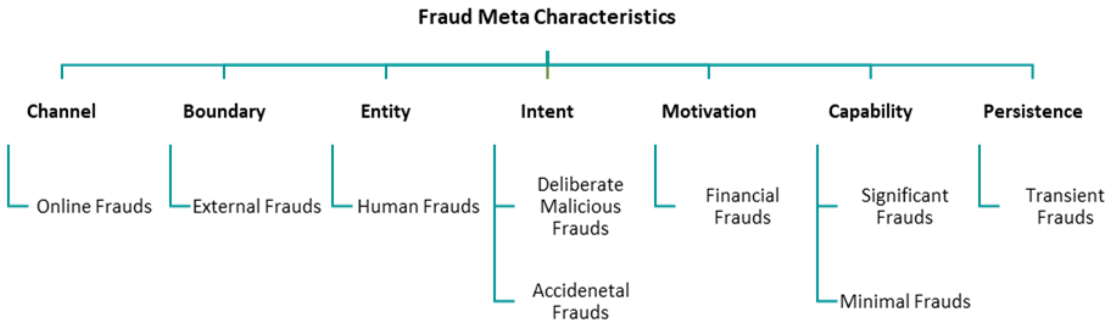

Figure 4.1 – Elementary fraud characteristics of this dissertation money mule case
Source: Adapted to the case analysed for the dissertation from Onwubiko (2020)

The money mule fraud process on the cryptocurrency exchange includes fiat deposits by newly joined users from a certain region; most profiles were between 18 and 30 years old. These are the characteristics of the confirmed cases provided by a third-party source. The pattern was specific because multiple small fiat deposits were exchanged for crypto, after which a more significant crypto withdrawal occurred. When first investigation described suspicious behaviour, there was no evident issue because the withdrawals were often sent to other exchanges or non-suspicious addresses. The frequency of the small fiat deposits was suspicious, followed by a more significant crypto withdrawal. These suspicions were then later confirmed by a third-party financial institution. The process of the money mule behaviour on the exchange is presented in Figure 4.2. The figure presents the process, starting when the user joins the exchange and concluding when the user makes a crypto withdrawal. There are other parts of this fraud, such as the subject laundering money and depositing the funds to the money mule account, but this occurred outside of the exchange, which means the selected company does not have specific information about what occurred before individuals joined the exchange.

In most cases, what happened to the crypto assets when the withdrawal was made is unclear. The issue is that the withdrawals were, in most cases, sent to exchanges or other service addresses. Tracking crypto assets through those addresses is very complicated because the frequency of transactions is high, and the assets are deposited to the exchange address, not individuals. The transactions are recorded on the ledger and then moved internally by the service. The transactions cannot be tracked publicly once the crypto is deposited to a service address, but the service can identify the subject that carried out the transaction (Chainalysis, 2020b).

For this dissertation, the set assumption was that what occurred after is not explicitly known, but that information is also irrelevant to the focus of the research. What was confirmed is the fact that these users were confirmed to be money mules, which was an essential piece of information for the research. I used those confirmed cases to build a model that will potentially predict which attributes determine money mules in order to prevent future fraud. The model is built upon those confirmed cases and is specific to the dataset but could potentially be applied to other cases.

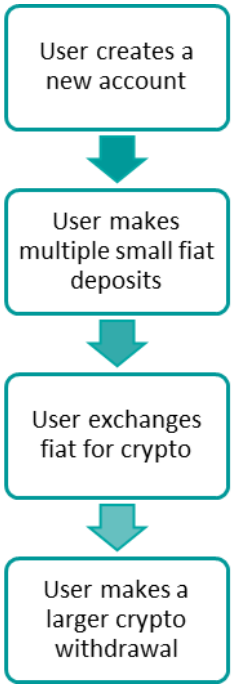

Figure 4.2 – Money mule fraud process on the cryptocurrency exchange

Source: Own work

4.2. Stage 2 – Data Understanding

The second stage of the CRISP-DM approach starts with the first steps concerning analysing the data. In the first stage, Business Understanding, I went through why the issue of fraud detection is essential in the company aspects. Having a basic understanding and goals of the analyses, I could proceed with dealing with the data. Firstly, I went through the dataset to learn the structure and composition of variables. That was followed by data preparation, which proved to be the most time-consuming part of the research (described in 4.3 Stage 3 – Data Preparation).

4.2.1. Data Collection and Description

The dataset originates from a cryptocurrency exchange and consists of users that joined during a specific period when the rise in money mule cases in a specific country was noticed. The data was collected from two sources, the general export from the company's user database and a list of confirmed money mule cases from the company's fraud department that gained the confirmed cases from a reliable third-party source.

Initially, when exporting the users for that period, there were around 22,000 records. However, a large number of those users had no transactions and, as such, did not fall into the scope of research, so they were removed from the dataset. Removing users resulted in a dataset of 6,384 users. The original data included 73 variables that consisted of numerical and categorical variables.

The initial data overview highlighted several issues:

- Removing around 16,000 users due to no transactions – initially cutting down from around 22,000 users for the investigated period to 6,384.

- Three users were missing from the dataset. Further research showed this was due to wrong input in the raw data. I added them manually during Stage 3 – Data Preparation. This step resulted in the dataset having 6,387 users.
- There is a significant data imbalance (class imbalance – when there are few cases of fraud and subsequently little data to identify them – one of the main challenges of fraud detection – (Guo et al. 2017)) of the dataset, having around 6387 cases and 123 confirmed cases, only 1.926% cases being fraudulent. The share of fraudulent cases is presented in Figure 4.3: All records: 6387
    - Confirmed money mules: 123 (1.926%)
    - Regular users: 6,264 (98.074%)

The list of confirmed cases for the period being researched included 123 users. The confirmed cases were added to the initial dataset and encoded as 1 for confirmed and 0 for non-fraudulent users.

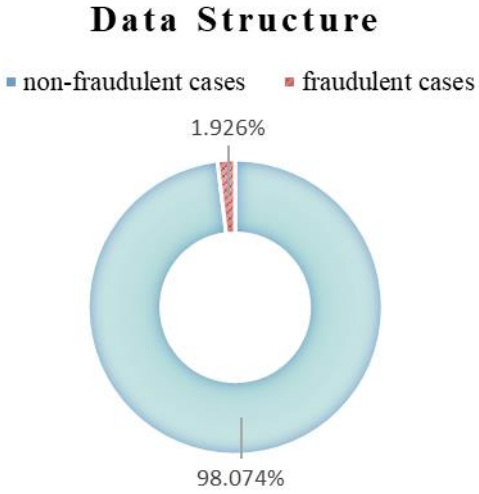

Figure 4.3 – Fraudulent and non-fraudulent user share – dataset structure
Source: Own work

4.2.2. Variables in the Initial Dataset

There were 73 columns in the initial dataset, many of which I then manipulated to achieve a more helpful format while also removing unnecessary information. Depending on the information theme, the main variables consist of personal information and other information about the user account, such as age, nationality, date and time of registration, and registration channel. Some information was gained through the exchange's blockchain analytical software, such as if the user had exposure to the dark web or if the crypto assets were deposited from a wallet or exchange. Additionally, the platform has a great deal of data related to the users' activities. Therefore, the variables used for the analysis include the types and sizes of deposits and withdrawals, averages, minimums and maximums, number of transactions and different crypto assets used.

4.2.3. Summary Statistics

The original dataset consisted of 73 columns, to which I added one column that indicated if the user is a confirmed money mule or not. From those 74 columns, 26 variables were saved as floating-point numbers, 26 as integers, and 22 as objects.

4.2.4. Descriptive Analytics

Within this subchapter is a basic outline of the original dataset before any modifications were made. Figure 4.4 presents the distribution of three categorical variables. There are three types of deposits that users use for the first deposit. The majority, almost 5,000 users, chose a fiat deposit, followed by credit cards and crypto deposits, representing less than 20%. Whereas fiat is the most popular deposit type, the same is not valid for withdrawals. More than 50% of all users in the dataset made the first withdrawal in crypto, overtaking fiat withdrawals by 5-10%.

Figure 4.4 – Distribution of categorical variables presented with histograms
Source: Own work

Subchapter 4.1.2 Money Mule Problem Description mentioned that the observed confirmed money mules made the first deposit in fiat and the first withdrawal in crypto. However, Figure 4.4 presents an overview of different categorical variables, where it can be observed that most users used the same methods for those same events. This means that only these two variables would probably not show the significance in the model as there is insufficient information.

Most of the users in the dataset that joined in the period used for analysis are in the age group between 18 (minimum sign-up age) to 35. Of those users, most carried out registration through the mobile application, whereas less than 20% carried it out through the website.

Figure 4.5 shows users' age distribution based on whether the user is a confirmed money mule. Looking at the figure, it is evident that most users fall into similar age dimensions for both scenarios. However, it is noticeable that the confirmed cases concentrate between 18 and 28 (35), whereas unconfirmed cases concentrate between 20 and 33 and go up to 50 if older outliers are disregarded. There are some outliers, mainly in the unconfirmed category, but also confirmed cases include a small number of users older than 35 – shown as small circles after the end of the graph whisker.

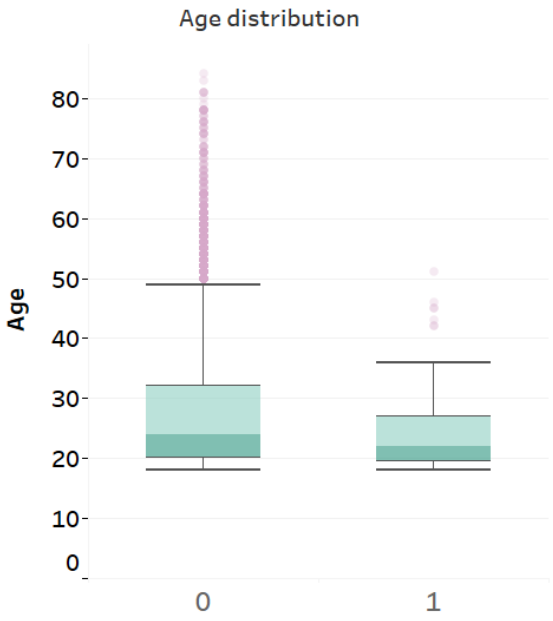

Figure 4.5 – Distribution of variable age

Source: Own work

Exploration of the data features of confirmed money mule cases presented some initial information about the cases. All the confirmed cases made a withdrawal. The majority first withdrew in BTC, ETH, or XRP, and some in fiat, but none used another crypto. Almost all users made their first deposit with fiat currencies. Information from the blockchain explorations showed that none of the confirmed cases received money from an exchange or had exposure to the dark market or fraud shop.

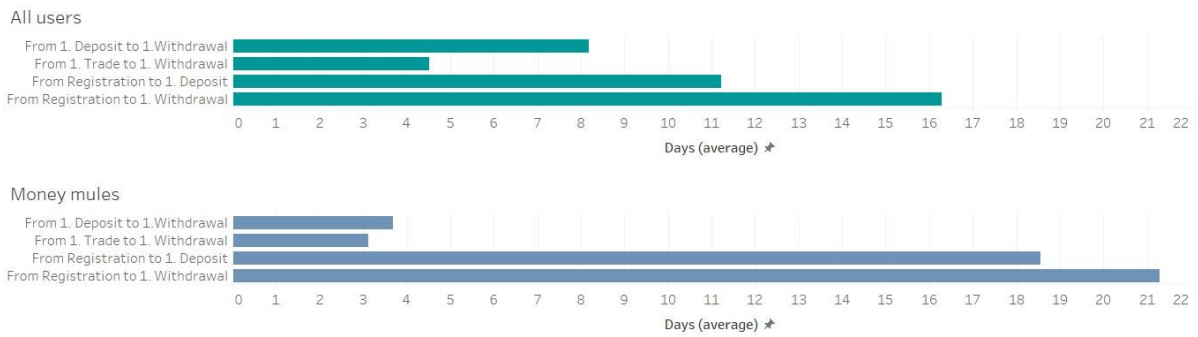

Figure 4.6 – Four variables showing the average days per category (money mules – bottom, all users – top)

Source: Own work

Figure 4.6 shows the average values for variables related to days between different actions. The top graph in Figure 4.6 shows data for all users, and the bottom graph shows the values for confirmed money mule cases. Confirmed money mules had a lower average for two variables: the time from first deposit to first withdrawal (1. Deposit to 1. Withdrawal) and first trade to first withdrawal (1. Trade to 1. Withdrawal). That means that those users performed those actions faster. They might have taken more time from registration to first deposit and consequently from registration to first withdrawal. However, once the deposit was made, all the consequent actions were performed faster than the average users.

4.2.5. Data Quality Issues

An initial overview of the dataset showed several columns with only 0 values. Some of the other columns in the dataset seemed irrelevant after further analysis. There were several columns with object types that can make analyses more complex. Most were objects unnecessarily being dealt with in the data preparation phase. Some of those were columns that had yes/no as strings. Another issue was the date variable, where a few rows had an incorrect data format.

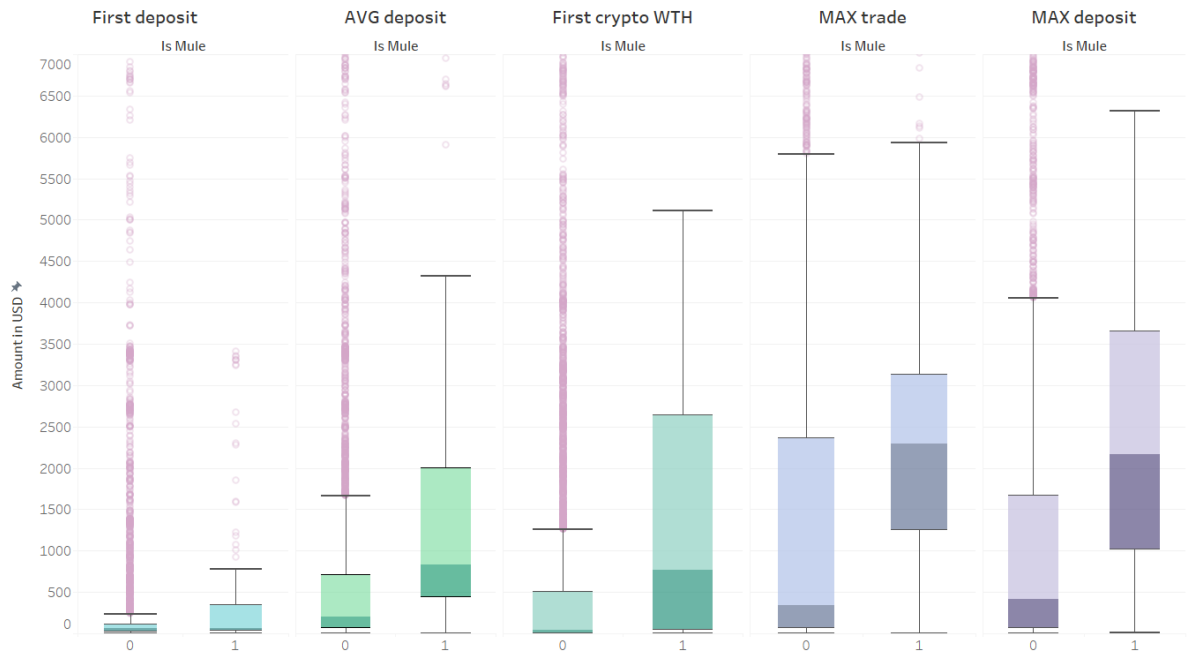

Figure 4.7 - Data distribution of several continuous variables

Source: Own work

Some other issues included the outliers in many continuous variables, visible in Figure 4.7. The graphs in the figure were limited on the y-axis depending on the variable, so the distribution can be seen, as the data included many more significant outliers. These variables include the size of the first deposit, average deposit size, first crypto withdrawal, maximum trade, and maximum deposit. The graphs in the figure show an evident difference between the medians of average deposit, maximum trade, and maximum deposit, depending on whether the users were money mules (marked as 0 and 1 at the

bottom of the graphs). That tells us that these variables could potentially be significant for the predictive models, but the outliers could affect the models' performance.

4.3. Stage 3 – Data Preparation

The second stage offered good insight into the dataset's characteristics and potential issues to be fixed. The third stage of the CRISP-DM is often the most demanding, as it focuses on dealing with issues in the dataset. Eliminating the issues and improving the data quality can have a distinct positive or negative effect on the precision of the model, depending on how well it is carried out (Chapman et al. 2000). I started the preparation by first adding the three missing users from the dataset and then focusing on more straightforward issues, such as duplicate and missing values and proceeding with outliers.

4.3.1. Removing Irrelevant Data and Dealing with Numerical Variables

An initial overview of the dataset showed several columns with only 0 values that would not affect our model, so I chose to remove them from the data. Other columns that were not relevant included "user_id," which was also removed for anonymisation purposes. I also removed the country of residence column, as the research only includes users from one country. However, I left the information of nationality. Several object variables had two options – yes/no, for example, if the money came to the user's account from an exchange. I changed those variables to an integer with values 1 and 0, which enabled the removal of some unnecessary object variables. The dataset included names of fraud shops or the dark web for users who carried out such transactions. I converted the two columns with names in strings to include only information if the user sent to those locations or not.

4.3.1.1. Handling Duplicate and Missing Values

I noticed that many variables had missing values through initial descriptive analytics. For most of those variables, I did not delete or replace those values. However, I did replace the missing values with 0 for the integer variables where the values are 0 or 1. This was carried out as the initial dataset had some fundamental data quality issues; one of them was missing 0 values. These variables include:

- Number of users who sent to the same withdrawal address
- Number of users who used the same credit card
- Number of trades before the first withdrawal
- Number of deposits before the first trade

4.3.2. Dealing with Categorical Variables

There were several categorical variables in the dataset, which I decided to one-hot encode. The process transforms categorical values into vector-based interpretation where each level of the categorical variable takes the value of 1 or 0 and is added as a new column in the dataset (Kelleher et al. 2020). Variables that were one-hot encoded were:

- Type of first withdrawal/deposit
- Joined with app/web

I added several columns related to the currencies. Original data included the following variables:

- Currency of first withdrawal
- Max traded currency
- Top pair traded

For those variables, I made two selections. I defined CORE crypto, including Bitcoin (BTC), Ethereum (ETH), and Ripple (XRP), and fiat currencies GBP, EUR, and USD. I then added dummy columns that indicate if the user's first deposit and the maximum traded currency were CORE crypto, fiat or other crypto assets. I added a variable for the top pair traded that indicates if the top trades included a CORE crypto asset.

4.3.3. Adding New Attributes

I added several columns to the dataset. Two were related to the acts carried out by users. Before, I mentioned that all the users left in the data frame had performed a deposit, because, otherwise, these users were irrelevant for our analysis. However, there were many missing values for certain variables from which I could deduce that not all users have made a trade or withdrawal. Therefore, I added two variables: one indicated if a user has made a trade and one if the user has made a withdrawal.

The data frame included one DateTime variable – the date and time the user joined the platform. From that, I added three variables: month, week, and day of the week the user joined.

Additionally, I added several variables depending on the age bin, as I could see that confirmed cases fell within a particular age group and wanted to test its significance on the model. Considering that most confirmed cases fell into the age range below 30, I added another variable that shows if the user is aged between 18 and 29.

4.3.4. Splitting the Dataset

The dataset was split into training and test parts. The training dataset consisted of 70% and the test of 30% of the original dataset as per similar research. The split was stratified to ensure that both the training and test sets have the same proportion of target variables because the dataset relates to a rare event with relatively few confirmed cases. Splitting the dataset into two independent sets ensures no information leakage, as the model is developed with the training set and then evaluated with the test set (Baesens et al. 2015).

The data split consisted of:

- Training dataset: 4470 instances (0 – 4384, 1 – 86)
- Training dataset: 1917 instances (0 – 1880, 1 – 37)

4.3.5. Standardisation

Some of the variables in the original dataset were continuous, such as the size of the first deposit, max trade, etc. These variables had a skewed distribution, as well as outliers. The dataset was standardised for logistic regression to make the models more reliable.

Standardisation is a feature scaling technique in which the standardised feature is calculated with Equation (11), where z is the standard score assigned to a value. x is the original feature value, μ is the mean of the x feature values, and σ is the standard deviation. Z-scores are values that show how many σ a value of a feature is away from the μ (Baesens et al., 2015).

$$z_i = \frac{x_i - \mu}{\sigma} \tag{11}$$

Source: Baesens et al. (2015)

After splitting the dataset into the test and training part, feature scaling was carried out. If the dataset was standardised before splitting, it could cause information leakage (Kelleher et al., 2020); therefore, I scaled the data separately for the training and test datasets and under and oversampling datasets. The data was scaled only for the logistic regression, as decision trees are not sensitive to the variance in the data. This is because the node split is not influenced by all features, but it occurs on a single feature. Since ensemble methods build models with decision trees, the feature scaling was unnecessary (Kelleher et al., 2020).

4.4. Stage 4 – Modelling

Five different algorithms were used to develop the predictive models. The modelling was done first with logistic regression and then decision trees. The models were not expected to perform that well because they work best on balanced data, which was not the case for the data in this dissertation. However, it was integral to use decision trees, as they are the building block of all the used ensemble methods. It was most useful to see how these models work for easier implementation in more complex models.

4.4.1. Feature Selection

The random forest feature importance approach was used to analyse feature importance. The results of the random forest being fit to the data are shown in Figure 4.8. The calculated importance served as a reference when assigning variables for the models. Some high-value importance features were slightly similar, so alternatives were chosen to reach peak model performance.

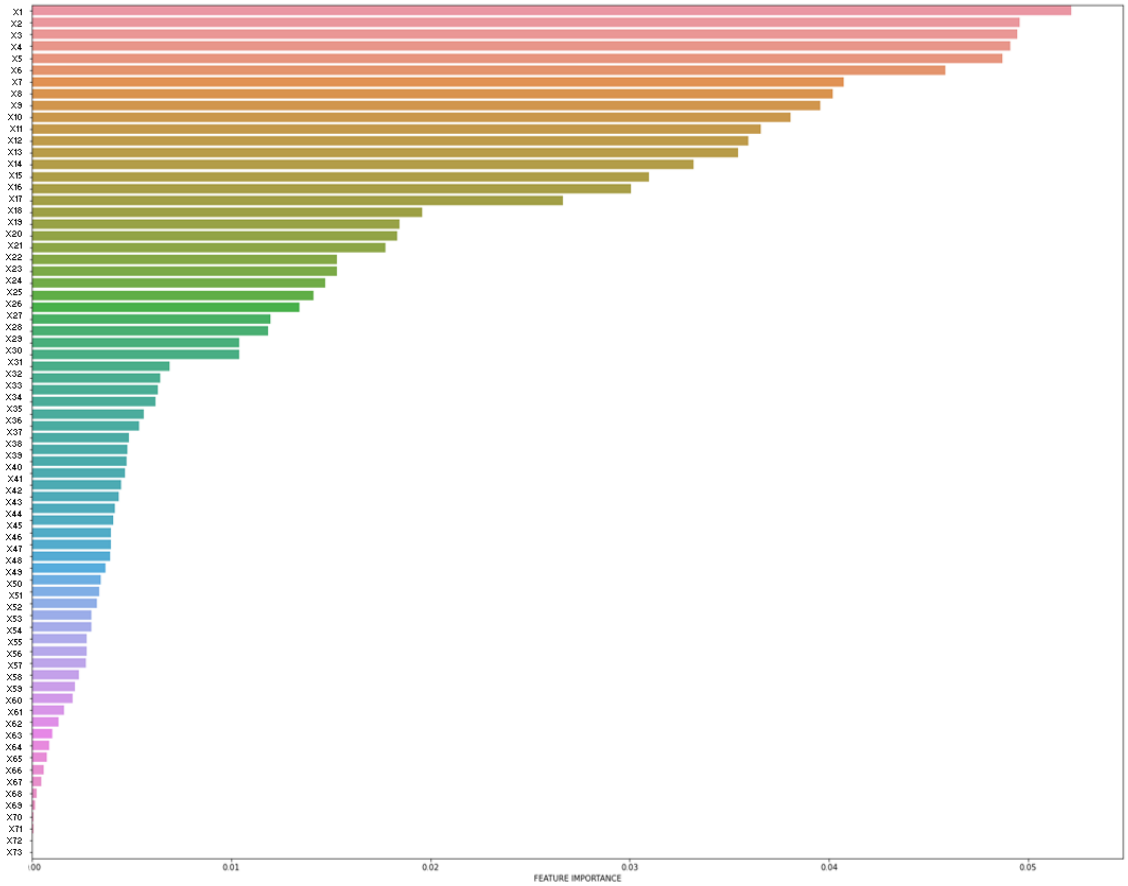

Figure 4.8 – Random forest *feature importance*
Source. Own work

4.4.2. Hyperparameter Optimisation

When setting up the models in Python, they take default hyperparameter values. For the models to perform as efficiently as possible, a grid search was run to test model performance and return optimal values for the hyperparameters. Optimised parameters differ depending on the implemented algorithm (Bentejac, Csorgo, and Martinez-Munoz, 2021). A 5-fold grid search was used when optimising the parameters, while the scoring method maximised AUC. The hyperparameters were optimised separately for the normal sample split and then separately for both sampling techniques. The grid search is applied to the training dataset to evaluate the most fitting parameters, which change with the change of the dataset. The parameters for models are outlined in Table 4.1.

Each algorithm has hyperparameters that can be optimised; some are the same, as the ensemble models contain decision trees, and some are different. The number of parameters is rather large, so the explanation for the parameters used only includes the ones I optimised in developed models. Most relevant hyperparameters that can be optimised to improve model performance of random forest include (Bentejac et al., 2021):

- max_features – the maximum number of features considered.
- min_samples_split – limits the tree size. Sets the minimum number of samples that are needed for an internal node split.

- min_samples_leaf – minimum number of samples that are needed to create a leaf/termination node.
- max_depth – maximum depth of a tree. This parameter is essential when trying to limit the overfitting of the model. Limiting the cut-off with lower depth could solve the issue if the model is overfitting.
- class_weight – the parameter that helps when data is imbalanced and improves the weights.

The most relevant hyperparameters that can be optimised to improve model performance of XGBoost and LightGBM include (Bentejac et al., 2021; XGBoost developers, n.d.):

- num_leaves (only LightGBM) – maximum number of tree leaves. The parameter helps control the tree complexity (like max depth in the decision tree and random forest)
- gamma (only XGBoost) – minimum loss reduction – the lower the value the deeper the tree
- max_depth
- scale_pos_weight – parameter for setting the balance between minority and majority cases – used for imbalanced data to improve representation.

To simplify the result description, the model names are abbreviated as:

- LR – Logistic regression
- DT – Decision tree
- RF – Random forest
- XGB – XGBoost
- LGBM - LightGBM

The sampling methods are abbreviated as:

- Train and test are the split training and test dataset without resampling.
- Under_0 is the training dataset and under_1 is the test dataset resampled with undersampling.
- Over_0 is the training dataset and over_1 is the test dataset resampled with oversampling.

Table 4.1 – Hyperparameter optimisation

Model	Hyper-parameter	Value
DT	max_depth	4
	min_samples_leaf	4
	min_samples_split	12
RF	max_depth	5
	min_samples_leaf	4
	min_samples_split	12
	class_weight	balanced
XGB	max_depth	2
	gamma	0.5
	subsample	0.6
	learining_rate	0.1

Model	Hyper-parameter	Value
LGBM	max_depth	2
	num_leaves	2
	learning_rate	0.25

Source: Own work

Optimisation of parameters was integral because when XGBoost was run, for example, the model was initially overfitted. This can be seen in Figure 4.9, where the AUC of the AUC_train, AUC_under_0, and AUC_over_0 were all almost 1, whereas the AUC when the models were tested were significantly lower. The results showed that the model performed incredibly well on the training dataset but could not be generalised well and performed poorly on the test dataset. With the optimisation of parameters, these results were improved.

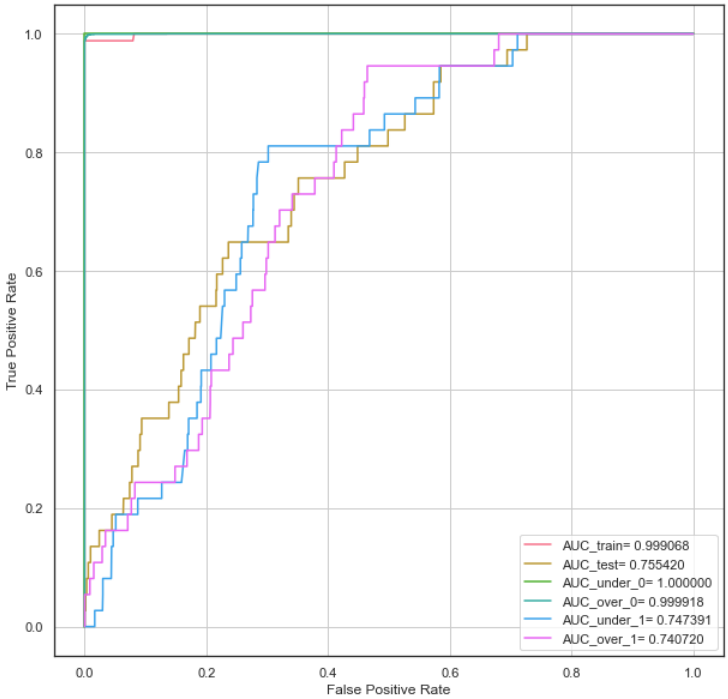

Figure 4.9 – XGBoost initial ROC curve

Source: Own work

For easier visualisation of results and understanding of the model, Figure 4.10 presents a decision tree extract from the random forest model. The figure shows the flow of tree development from the root node to the leaf/termination nodes.

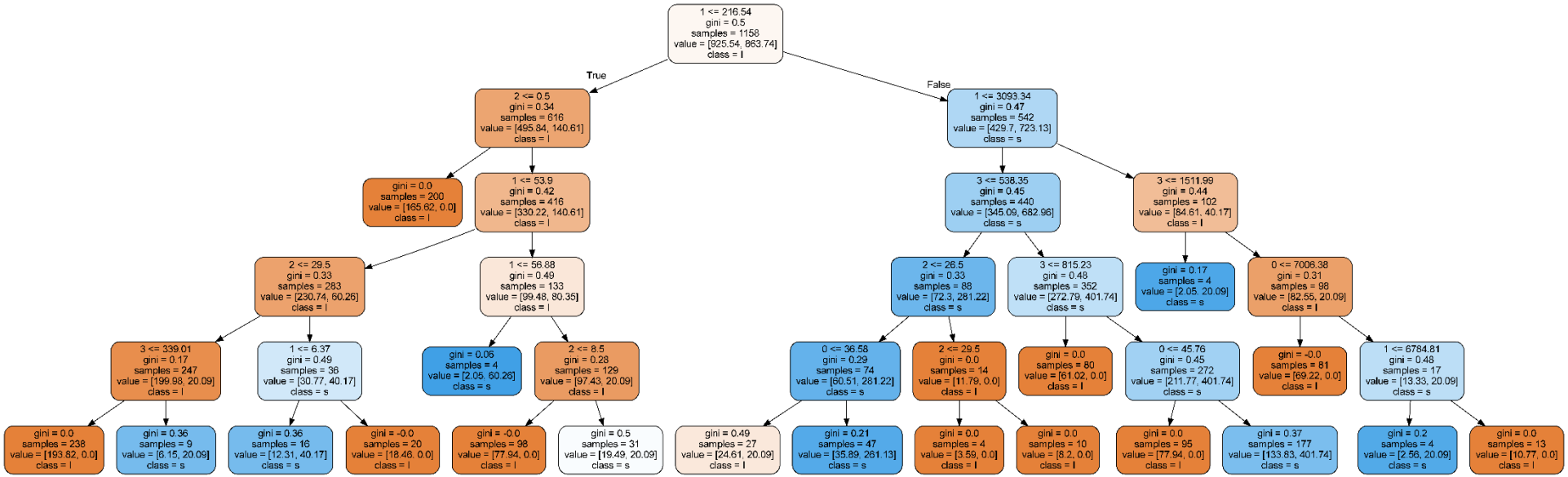

Figure 4.10 – Decision tree from the random forest

4.4.3. Model Results

After finishing the model optimisation and gaining results, the final values were extracted to prepare various figures to present the results. The complete table of value results is presented in Appendix 2 and consists of the matrices presented in 2.6.4 Metrics for Classification. The two main metrics are confusion matrix and AUC. It's important to note that a high AUC does not necessarily make a model great. The values in the confusion matrix and the calculated values such as TPR and FNR, as well as misclassification rate (MR) must also be considered.

The classification metrics are congregated in Figure 4.18, which includes the AUC and metrics calculated from confusion metrics results for each model. The complete prediction results are in Appendix 1. These results were then used to make other classification metrics, such as classification accuracy (CA), misclassification rate (MR), true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR). Classification results are partially presented in Table 4.2, which includes the prediction results for the models when assessed with the test dataset.

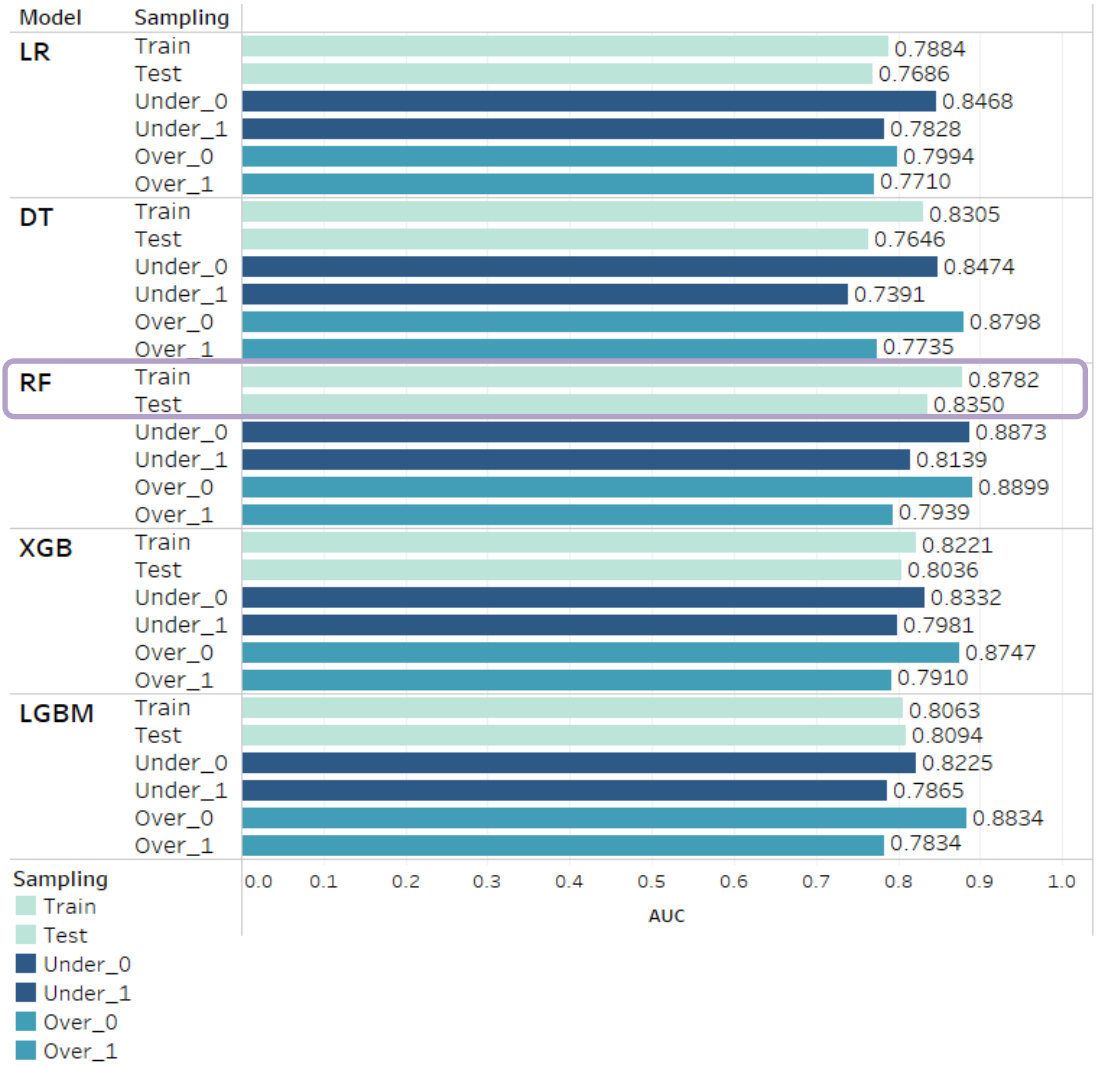

Figure 4.11 - AUC results for all models based on sampling

Source: Own work

Figure 4.11, 4.12 and 4.18 show a visual comparison of results depending on the model and sampling technique. Not only is it important for the test results to be good, but the consistency of going from training results to testing results is also imperative. The training results are marked as "Train" – without resampling, "Under_0" for Undersampling, and "Over_0" for oversampling. The test results are marked as "Test" for those without resampling, "Under_1" for undersampling, and "Over_1" for oversampling.

The model development was initially approached with LR and then a DT. The AUC values of all models based on the sampling technique are presented in Figure 4.11. Looking at individual model performance regarding AUC for LR, the best performing model is the one that was trained on an under-sampled dataset. However, when the model was tested, the MR (shown in Figure 4.14) was the highest of all tested LR models.

Figure 4.12 – Ensemble results of model testing

Source: Own work

The model developed with decision trees was initially overfitting, which meant the tree growth needed to be limited and hyperparameters optimised. The results after applying those changes were encouraging, as they showed an improvement from the LR.

With a basic understanding of tree algorithms and basic model development, the next step was applying ensemble methods using trees. Using the RF was the next progression. The results were again an improvement from the DT. For the RF, all models' AUC metrics results ranked higher than the LR and DT for all the sampling techniques.

For all the ensemble methods – RF, XGB, and LGBM – AUC results were best for models where the training dataset is the original and the imbalanced data issue was optimised with hyperparameters. These models also had the most stable performance regarding the difference between the training and testing dataset results.

The idea was to try more advanced ensemble methods, in addition to RF, to improve the results. However, it turned out that the best performing ensemble method, according to the AUC metrics, was RF and not gradient boosting. The metrics for ensemble methods results when models were tested are visually presented in Figure 4.12. XGB and LGBM managed to predict more TP classifications, resulting in higher TPR or sensitivity for the basic and undersampling models. However, the TNR, or specificity, was lower for both boosting techniques, resulting in a higher number of false positive predictions. Generally, sensitivity is of high importance for fraud predictions, but maximising that worsened the AUC results substantially and made specificity unremarkable.

Table 4.2 – Classification results for models based on sampling technique

Model	Sampling	TP	TN	FP	FN
LR	Test	27	1283	597	10
	Under_1	31	1165	715	6
	Over_1	26	1288	592	11
DT	Test	33	1123	757	4
	Under_1	24	1393	487	13
	Over_1	23	1327	553	14
RF	Test	30	1329	551	7
	Under_1	33	1154	726	4
	Over_1	23	1382	498	14
XGB	Test	31	1255	625	6
	Under_1	34	1109	771	3
	Over_1	23	1380	500	14
LGBM	Test	34	1131	749	3
	Under_1	34	1091	789	3
	Over_1	23	1386	494	14

Source: Own work

The predictive models developed with XGB and LGBM showed good results, but these ensemble methods at a high risk for overfitting (Kelleher et al., 2020), This was also shown in Figure 4.9, , which shows the initial XGB ROC curve. The overfitting was rectified through hyperparameter optimisation, but the end AUC scores were lower than RF.

The random forest model trained on the original dataset split without resampling showed the most promising results because the AUC score is the highest while keeping the FNR and FPR low. Some XGB

and LGBM models offered better TPR results, resulting in lower TNR, meaning that the RF model performs better than boosting.

RF ROC curve model is shown in Figure 4.15, where the model with an AUC value of 0.83 is indicated by an orange line. The graph shows that the training model AUC values for under and oversampling are higher than the original data. However, the models were not as stable and generalisable as the basic ones because when the model was tested on the test dataset, the models performed worse than the model trained without resampling.

Figure 4.13 shows the performance of LG, where performance is best with Undersampling, but the results are slightly better than DT, shown in Figure 4.14, where there were also issues with overfitting.

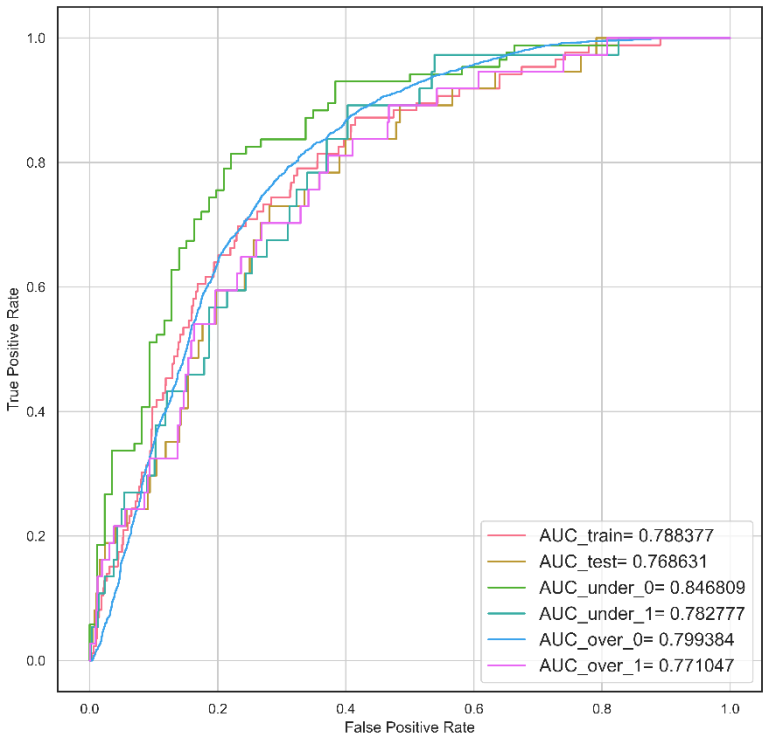

Figure 4.13 - Logistic regression - ROC curve

Source: Own work

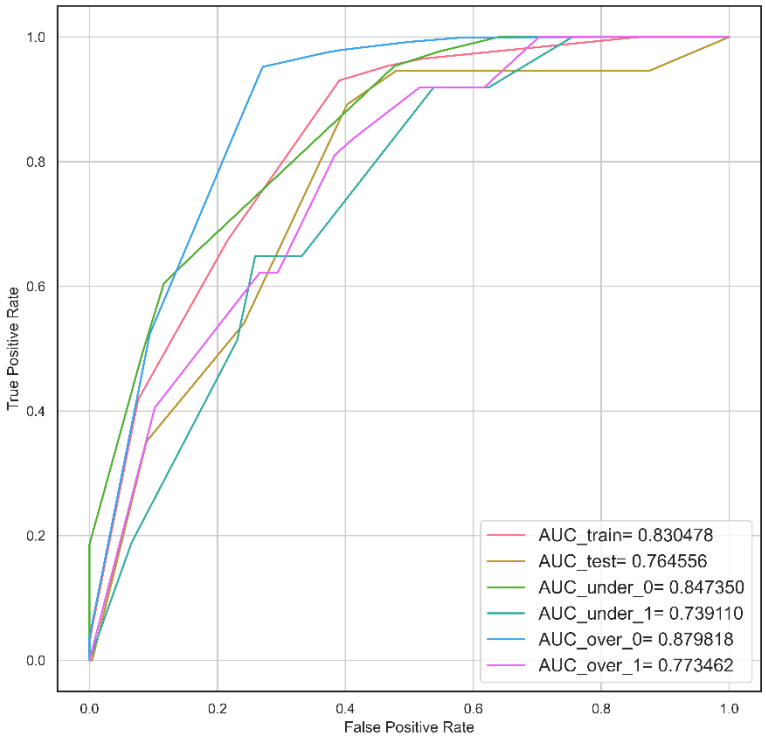

Figure 4.14 – Decision tree ROC curve

Source: Own work

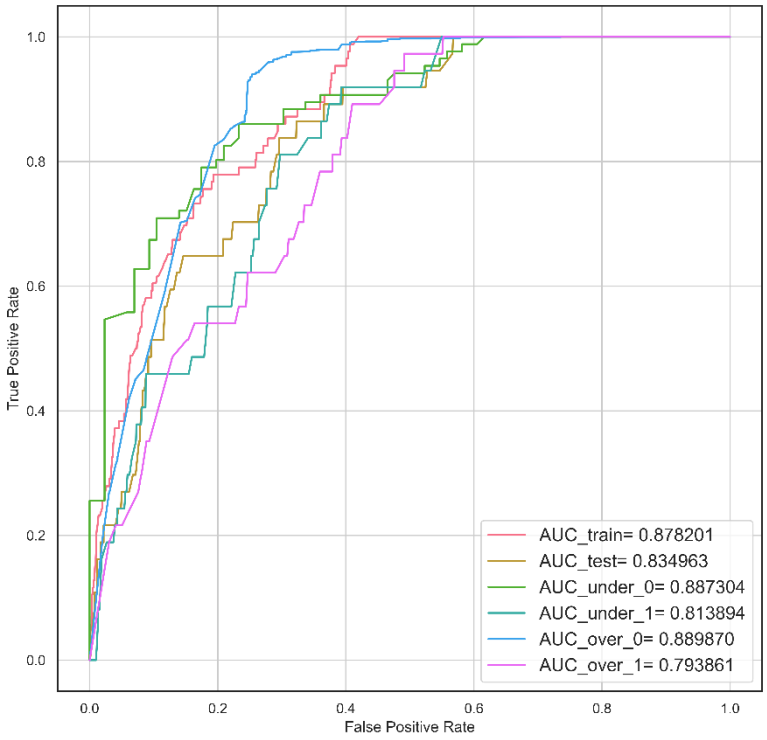

Figure 4.15 – Random forest ROC curve

Source: Own work

Figure 4.16 and Figure 4.17 show ROC curves for boosting algorithms, XGB and LGBM. Both models perform better without resampling methods

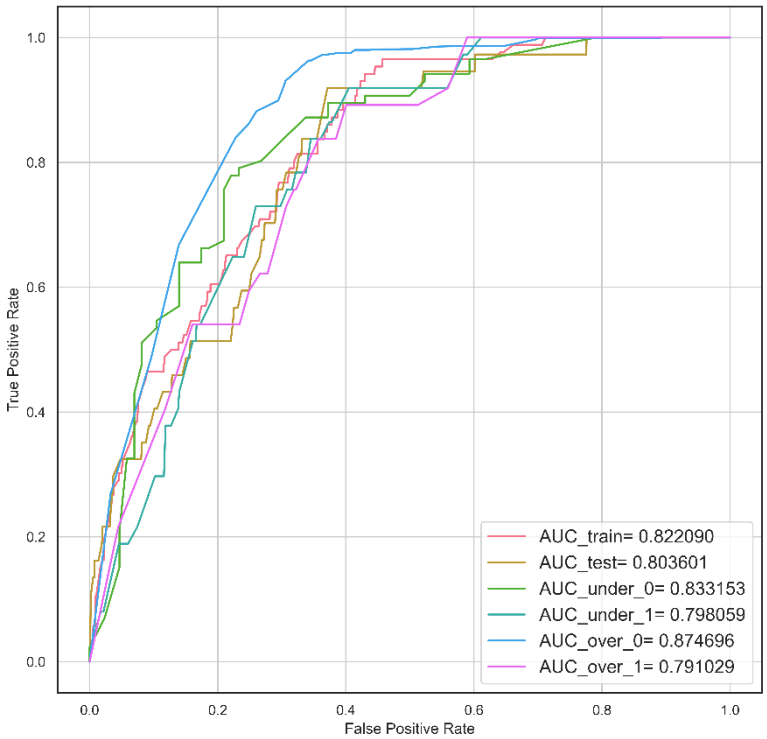

Figure 4.16 – XGBoost ROC curve

Source: Own work

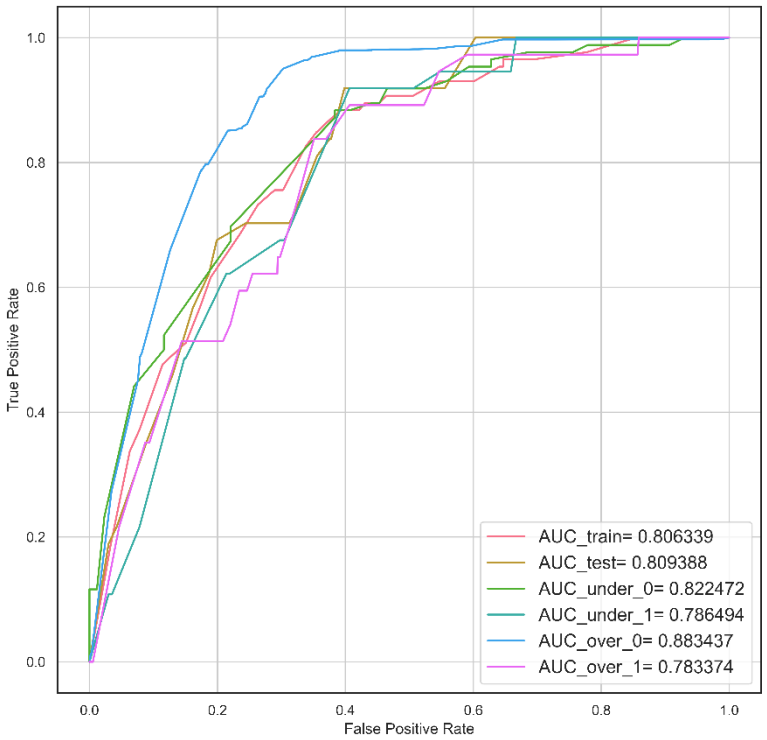

Figure 4.17 – LightGBM ROC curve

Source: Own work

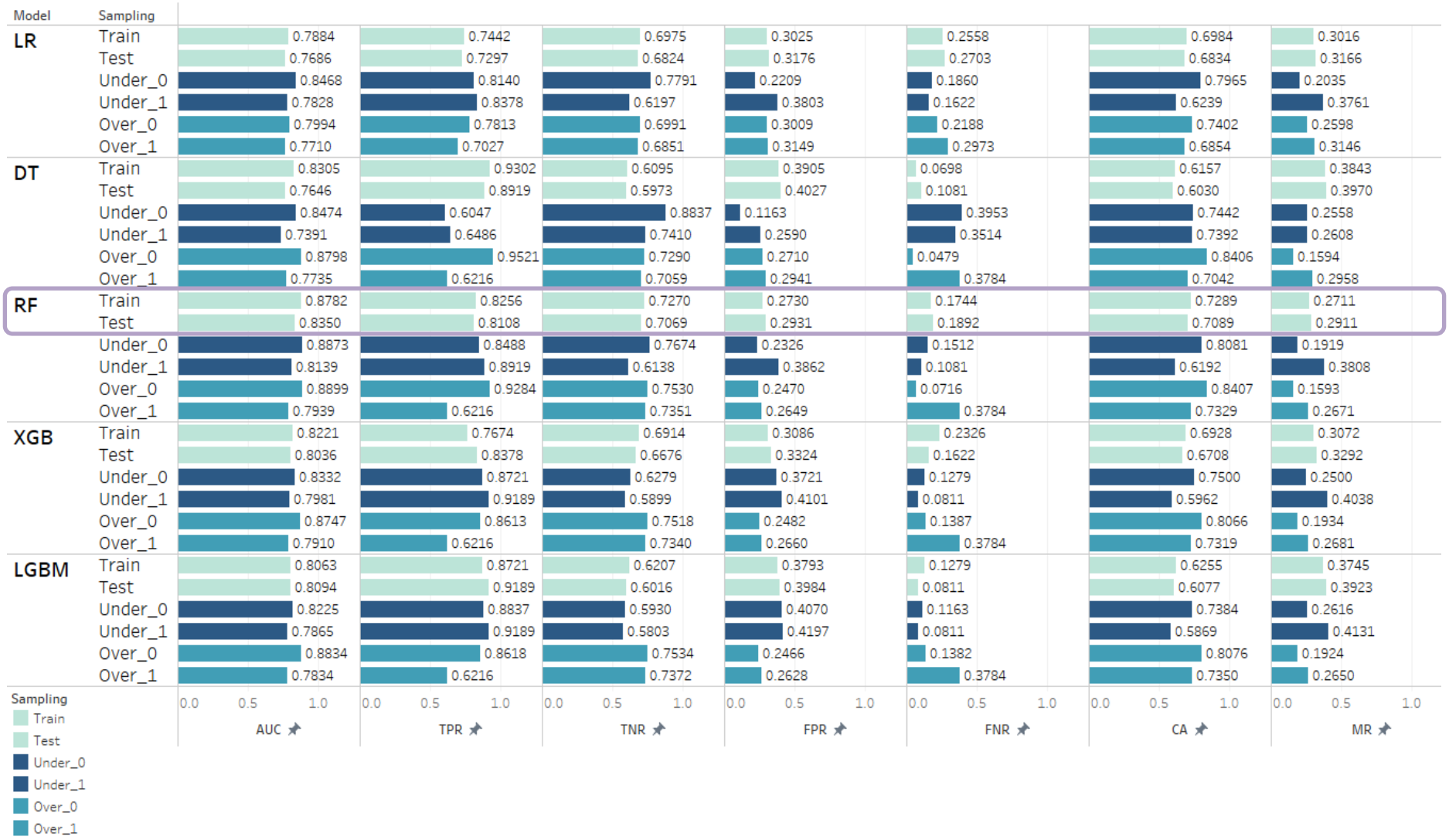

Figure 4.18 - Results of all metrics

Source: Own work

4.5. Stage 5 – Evaluation

The predictive models developed with several machine learning algorithms perform adequately in terms of chosen metrics. The results of the predictive analytics models are, according to AUC values, rather good, taking values above 0.8. However, the situation is not ideal when looking at false positive and false negative cases. The number of false predictions could cause substantial strains on company resources. Generally, the model results are good, but when compared to other similar analyses done with credit card fraud, the results are not as nearly efficient. Baesens, Vlasselaer, and Verbeke (2015) and West and Bhattacharya (2016) all report performance above 0.9 AUC values. That is not necessarily a sign of poor performance for the issue because credit card fraud differs from money mule cases in the cryptocurrency industry. Major differences include data availability and quantity because the money mule cases were very rare in a small dataset. However, not much research has been done in this area, so there are not many benchmark results to compare.

The business objectives were presented in 4.2 Stage 2 – Data Understanding and set out to find a solution for the money mule issue the selected company encountered. The idea was to develop a model to help minimise the risk to the business that these fraudulent users present. The models, if implemented, could facilitate the optimisation of fraud detection. The best performing model is the random forest, which detected 81% of all positive cases. The model classified 29% of negative cases as positive. Those are the two main issues: 19% incorrectly classified positive cases and 29% incorrectly classified negative cases. The money mule prediction model has weaker performance when compared to literature, where West and Bhattacharya (2016) report accuracies above 95%. The implementation would improve the existing processes but could potentially cause other business limitations. Part of the objective was to minimise risk to the company's reputation, which could be one of the drawbacks of implementing developed models. The models would minimise the risk of fines and elevation of risk scores, but classifying legitimate users incorrectly as money mules would negatively affect the selected company's reputation. An additional drawback would be increased workload for the fraud department and other stakeholders within the company, which could at least temporarily lower employee satisfaction before new processes would be implemented to support the increased workload. The potential implementation of the model and use of results variations are discussed in the following chapter about deployment.

Another option would be to use logistic regression output, the probability of an event being fraudulent or not. However, the model was not as well performing as the random forest, which means the implementation could result in additional drawbacks.

4.6. Stage 6 – Deployment

Deployment is the last stage of the CRISP-DM methodology, where we consider next steps for the models. The previous stage, Stage 5 – Evaluation, gave an overview of how the developed models work, considering the business goals set out at the beginning of the data analysis. As presented, the results are not the most optimal when considering how to ensure that implementation would not cause more constraints than benefits. That means that the models will most likely not be implemented directly.

Therefore, the plan for the future of the models is to present them to the stakeholders in the selected company and review them for further improvements. There are multiple options as to what could happen to the developed models:

- The models are further improved when there is more available data. The data will still be imbalanced because money mules are rare events, and predicting rare events is challenging (Guo et al., 2017). The analysis could be expanded to not only one country but perhaps by region, as there might be some patterns in larger data. One of the constraints was the amount of data, as there were not many confirmed cases of money mules. The company could also add more KYC information from the users, providing additional information that could potentially be relevant for making predictions.
- A dashboard is designed to serve as a reference for employees of the selected company that carry out processes related to money mule behaviour. The models would not be implemented due to suboptimal performance but could potentially improve some workflows.

As mentioned, the models most likely will not be directly implemented into the selected company's processes. However, if it were, other processes would need to be implemented. It is possible that the selected company could decide to implement the money mule detection solution. However, as there would be many false positive cases, these occurrences need to be investigated due to a suspicion of illicit behaviour. However, to implement a solution like that, additional technological solutions would have to be implemented to avoid too many process bottlenecks. Employees checking the illicit behaviour would need assistance to minimise the constraints on the department. An additional issue would be the customer service aspect. A new, user-friendly system could be implemented that would automate the check for essential documents and allow users to confirm their identity and disprove fraudulent activity rather easily and quickly. That would help minimise user churn due to complex processes while ensuring the company complies with all regulations and helps minimise money laundering that flows through the cryptocurrency exchange.

5. CONCLUSIONS

This dissertation addressed an issue with real-life data that originated from a cryptocurrency exchange. The goal to develop predictive analytics models with supervised machine learning algorithms was reached by successfully developed models of logistic regression, decision tree, and ensemble methods – random forest, XGBoost, and LightGBM. The model performance was optimised by trying oversampling, undersampling the dataset, and optimising hyperparameters for individual algorithms. The selected algorithms and techniques showed encouraging results in existing studies for fraud detection when the dataset is highly imbalanced, which was the case for the dataset used in this dissertation. The model development resulted in several promising predictive models optimised to find the most fitting one for the issue that transpired at the selected company – a centralised cryptocurrency exchange. The model where the classification metrics showed results best aligned with business goals was developed with random forest without resampling the dataset.

The type of money laundering analysed seemed to be a challenging issue for predictive analytics and not trivial. The reality is that the confirmed fraud cases perhaps did not have features specific enough to be distinguishable, as the data was highly imbalanced, with only 1.92% classified as money mules. Money mule cases are incredibly hard to detect, especially when cryptocurrencies that offer fast transactions and pseudo-anonymity are involved. The individuals involved do not perform actions that would trigger thresholds, whereas the whole process is carried out relatively fast. It could be interesting that even if money mules cannot be stopped before transferring the assets, they could be identified as money mules and the assets tracked on the blockchain. Since money mules are often not compliant but unwitting or unknowing, it would potentially be of interest to the authorities to improve the identification of the criminals carrying out the money laundering activities.

The purpose of the dissertation, to help the selected company minimise the occurrence of money mules and mitigate the risk associated with it, was perhaps not fully reached thus far. The business objectives are partially met with a model that predicts most fraudulent cases. The issue, however, is that the model also falsely predicts some classifications. The problem is twofold; firstly, the model does not predict all positive cases, and secondly, the number of false positive cases is rather substantial. The first part means that the exchange does not fully avoid the risk of money mule related behaviour, which still presents an AML risk to the selected company. The second part of the issue, false positive cases, could cause constraints for business resources, as the potentially positive cases would need to be investigated. To investigate that high percentage of cases would mean many resources would need to be utilised, potentially causing bottlenecks in fraud processes. Additionally, falsely accusing clients of carrying out illicit behaviour would result in damage to the company's image. The latter could be avoided by automating some of the processes. Even if the model is not 100% accurate, the false positive cases can be dealt with using an automated initial process for the clients to securely provide proof of innocence. That would result in the selected company at least partially minimising the risk of illicit behaviour with fewer resources than initially projected.

The results show potential and already provide some business value, but additional research should be done with more available data to impact business operations more wholesomely. The selected company can use the results and model as a reference tool to confer on and as a base for further data analyses.

The developed model performance shows that the predictive analytics approach to fraud detection also has some potential for application in the cryptocurrency industry. What needs further research is that, contrary to older financial institutions, cryptocurrency exchanges operate faster and have several complex challenges, such as higher anonymity of transactions. Money mules are an instrument criminals use to hide the origins of their wealth. This type of money laundering is often hard to detect because the transactions are carried out by individuals who previously did not show suspicious behaviour and are below thresholds. For the fraud detection predictive models to perform better, there needs to be better data availability in terms of more thorough KYC processes and more widespread quality compliance practices within the whole industry.

6. BIBLIOGRAPHY

Alsuwailem, Alhanouf Abdulrahman Saleh, and Abdul Khader Jilani Saudagar. 2020. 'Anti-Money Laundering Systems: A Systematic Literature Review'. *Journal of Money Laundering Control* 23(4):833–48. doi: 10.1108/JMLC-02-2020-0018.

Antonopoulos, Andreas M. 2017. *Mastering Bitcoin*. O'Reilly Media, Inc.

Arslanian, Henri, and Fabrice Fischer. 2019. *The Future of Finance: The Impact of FinTech, AI, and Crypto on Financial Services*. Cham: Springer International Publishing.

Baesens, Bart, Veronique Van Vlasselaer, and Wouter Verbeke. 2015. *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*. 1st ed. Wiley Publishing.

Beck, Adam. 2002. 'Hashcash - A Denial of Service Counter-Measure'.

Bentejac, C., A. Csorgo, and G. Martinez-Munoz. 2021. 'A Comparative Analysis of Gradient Boosting Algorithms'. *ARTIFICIAL INTELLIGENCE REVIEW* 54(3):1937–67. doi: 10.1007/s10462-020-09896-5.

Berwick, Angus, and Tom Wilson. 2022. 'Crypto Giant Binance Kept Weak Money-Laundering Checks, Documents Show'. *Reuters*.

Campbell-Verduyn, Malcolm. 2018. 'Bitcoin, Crypto-Coins, and Global Anti-Money Laundering Governance'. *Crime, Law and Social Change* 69(2):283–305. doi: http://dx.doi.org/10.1007/s10611-017-9756-5.

Chainalysis. 2020a. 'Chainalysis in Action: US Government Agencies Seize More Than $1 Billion in Cryptocurrency Connected to Infamous Darknet Market Silk Road'. *Chainalysis*. Retrieved 27 March 2022 (https://blog.chainalysis.com/reports/silk-road-doj-seizure-november-2020/).

Chainalysis. 2020b. 'Why You Can't Trace Funds Through Services Using Blockchain Analysis (And Why You Don't Need to Anyway)'. *Chainalysis*. Retrieved 11 June 2022 (https://blog.chainalysis.com/reports/blockchain-analysis-trace-through-service-exchange/).

Chainalysis. 2021. 'FATF's Updated Guidance Tells Regulators to Focus on Business Models Over Technology and Terminology'. *Chainalysis*. Retrieved 22 July 2022 (https://blog.chainalysis.com/reports/fatf-cryptocurrency-guidance-october-2021/).

Chainalysis. 2022. *The 2022 Crypto Crime Report*.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. 2000. 'CRISP-DM 1.0: Step-by-Step Data Mining Guide'. *SPSS Inc* 9:13.

Chen, Zhiyuan, Le Dinh Van Khoa, Ee Na Teoh, Amril Nazir, Ettikan Kandasamy Karuppiah, and Kim Sim Lam. 2018. 'Machine Learning Techniques for Anti-Money Laundering (AML) Solutions in Suspicious Transaction Detection: A Review'. *Knowledge and Information Systems* 57(2):245–85. doi: 10.1007/s10115-017-1144-z.

CoinMarketCap. 2022. 'Cryptocurrency Prices, Charts And Market Capitalizations'. *CoinMarketCap*. Retrieved 14 March 2022 (https://coinmarketcap.com/).

Dorminey, Jack, A. Scott Fleming, Mary-Jo Kranacher, and Richard A. Riley Jr. 2012. 'The Evolution of Fraud Theory'. *Issues in Accounting Education* 27(2):555–79. doi: 10.2308/iace-50131.

Esoimeme, Ehi Eric. 2021. 'Identifying and Reducing the Money Laundering Risks Posed by Individuals Who Have Been Unknowingly Recruited as Money Rules'. *Journal of Money Laundering Control* 24(1):201–12. doi: http://dx.doi.org/10.1108/JMLC-05-2020-0053.

Europol. 2021. *Cryptocurrencies - Tracing the Evolution of Criminal Finances*. Luxembourg: Publications Office of the European Union.

EUROPOL. 2021. 'Money Muling'. *Europol*. Retrieved 9 May 2022 (https://www.europol.europa.eu/operations-services-and-innovation/public-awareness-and-prevention-guides/money-muling).

FATF. 2021. *Updated Guidance for a Risk-Based Approach to Virtual Assets and Virtual Asset Service Providers*. Paris: FATF.

Federal Bureau of Investigation. n.d. 'Money Mules'. *Federal Bureau of Investigation*. Retrieved 21 July 2022 (https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/money-mules).

Fletcher, Emily, Charles Larkin, and Shaen Corbet. 2021. 'Countering Money Laundering and Terrorist Financing: A Case for Bitcoin Regulation'. *Research in International Business and Finance* 56:101387. doi: 10.1016/j.ribaf.2021.101387.

Fosso Wamba, Samuel, Jean Robert Kala Kamdjoug, Ransome Epie Bawack, and John G. Keogh. 2020. 'Bitcoin, Blockchain and Fintech: A Systematic Review and Case Studies in the Supply Chain'. *Production Planning & Control* 31(2–3):115–42. doi: 10.1080/09537287.2019.1631460.

Gandomi, Amir, and Murtaza Haider. 2015. 'Beyond the Hype: Big Data Concepts, Methods, and Analytics'. *International Journal of Information Management* 35(2):137–44. doi: 10.1016/j.ijinfomgt.2014.10.007.

Guo, HX, YJ Li, J. Shang, MY Gu, YY Huang, and G. Bing. 2017. 'Learning from Class-Imbalanced Data: Review of Methods and Applications'. *EXPERT SYSTEMS WITH APPLICATIONS* 73:220–39. doi: 10.1016/j.eswa.2016.12.035.

Haber, Stuart, and W. Scott Stornetta. 1991. 'How to Time-Stamp a Digital Document'. *Journal of Cryptology* (3):99–111.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. edited by T. Hastie, R. Tibshirani, and J. Friedman. New York, NY: Springer.

Kelleher, John, Brian Mac Namee, and Aoife D'Arcy. 2020. *Fundamentals of Machine Learning for Predictive Data Analytics : Algorithms, Worked Examples, and Case Studies*. Second edition. Cambridge, Massachusetts: The MIT Press.

Lu, Yang. 2018. 'Blockchain and the Related Issues: A Review of Current Research Topics'. *Journal of Management Analytics* 5(4):231–55. doi: 10.1080/23270012.2018.1516523.

Microsoft. n.d. 'Welcome to LightGBM's Documentation! — LightGBM 3.3.2.99 Documentation'. Retrieved 8 August 2022 (https://lightgbm.readthedocs.io/en/latest/Features.html).

Nakamoto, Satoshi. 2008. 'Bitcoin: A Peer-to-Peer Electronic Cash System'.

Ngai, E. W. T., Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. 2011. 'The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature'. *On Quantitative Methods for Detection of Financial Fraud* 50(3):559–69. doi: 10.1016/j.dss.2010.08.006.

Onwubiko, Cyril. 2020. 'Fraud Matrix: A Morphological and Analysis-Based Classification and Taxonomy of Fraud'. *Computers & Security* 96:101900. doi: 10.1016/j.cose.2020.101900.

Provost, Foster, and Tom Fawcett. 2013. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. 1st ed. O'Reilly Media, Inc.

Salehi, Ahmad, Mehdi Ghazanfari, and Mohammed Fathian. 2017. 'Data Mining Techniques for Anti Money Laundering'. 12(20):11.

TRM. n.d. 'Transaction Monitoring'. Retrieved 13 August 2022 (https://www.trmlabs.com/products/transaction-monitoring).

United Nations. n.d. 'Money Laundering'. *United Nations : Office on Drugs and Crime*. Retrieved 22 July 2022 (//www.unodc.org/unodc/en/money-laundering/overview.html).

W. Diffie and M. Hellman. 1976. 'New Directions in Cryptography'. *IEEE Transactions on Information Theory* 22(6):644–54. doi: 10.1109/TIT.1976.1055638.

West, Jarrod, and Maumita Bhattacharya. 2016. 'Intelligent Financial Fraud Detection: A Comprehensive Review'. *Computers & Security* 57:47–66. doi: 10.1016/j.cose.2015.09.005.

Wolfe, David T., and Dana R. Hermanson. 2004. 'The Fraud Diamond: Considering the Four Elements of Fraud'. *CPA Journal* 74(12):38–42.

XGBoost developers. n.d. 'XGBoost Documentation — Xgboost 1.5.2 Documentation'. Retrieved 19 March 2022 (https://xgboost.readthedocs.io/en/stable/).

Zhang, Chongsheng, Changchang Liu, Xiangliang Zhang, and George Almpanidis. 2017. 'An Up-to-Date Comparison of State-of-the-Art Classification Algorithms'. *Expert Systems with Applications* 82:128–50. doi: 10.1016/j.eswa.2017.04.003.

Zhang, Yan, and Peter Trubey. 2019. 'Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection'. *Computational Economics* 54(3):1043–63. doi: 10.1007/s10614-018-9864-z.

Zheng, Z., S. Xie, H. Dai, X. Chen, and H. Wang. 2017. 'An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends'. Pp. 557–64 in *2017 IEEE International Congress on Big Data (BigData Congress)*.

7. APPENDIX

Appendix 1 – Confusion matrix results for all models

Model	Sampling	TP	TN	FP	FN
LR	Train	64	3058	1326	22
	Test	27	1283	597	10
	Under_0	70	67	19	16
	Under_1	31	1165	715	6
	Over_0	3425	3065	1319	959
	Over_1	26	1288	592	11
DT	Train	80	2672	1712	6
	Test	33	1123	757	4
	Under_0	52	76	10	34
	Under_1	24	1393	487	13
	Over_0	4174	3196	1188	210
	Over_1	23	1327	553	14
RF	Train	71	3187	1197	15
	Test	30	1329	551	7
	Under_0	73	66	20	13
	Under_1	33	1154	726	4
	Over_0	4070	3301	1083	314
	Over_1	23	1382	498	14
XGB	Train	66	3031	1353	20
	Test	31	1255	625	6
	Under_0	75	54	11	32
	Under_1	34	1109	771	3
	Over_0	3776	3296	1088	608
	Over_1	23	1380	500	14
LGBM	Train	75	2721	1663	11
	Test	34	1131	749	3
	Under_0	76	51	35	10
	Under_1	34	1091	789	3
	Over_0	3778	3303	1081	606
	Over_1	23	1386	494	14

Source: Own work

Appendix 2 – Classification metrics results for all the models

Model	Sampling	TPR	TNR	FPR	FNR	MR	CA	AUC
LR	Train	0.744186	0.697536	0.302464	0.255814	0.301566	0.698434	0.788377
LR	Test	0.729730	0.682447	0.317553	0.270270	0.316641	0.683359	0.768631
LR	Under_0	0.813953	0.779070	0.220930	0.186047	0.203488	0.796512	0.846809
LR	Under_1	0.837838	0.619681	0.380319	0.162162	0.376109	0.623891	0.782777
LR	Over_0	0.781250	0.699133	0.300867	0.218750	0.259808	0.740192	0.799384
LR	Over_1	0.702703	0.685106	0.314894	0.297297	0.314554	0.685446	0.771047
DT	Train	0.930233	0.609489	0.390511	0.069767	0.384340	0.615660	0.830478
DT	Test	0.891892	0.597340	0.402660	0.108108	0.396974	0.603026	0.764556
DT	Under_0	0.604651	0.883721	0.116279	0.395349	0.255814	0.744186	0.847350
DT	Under_1	0.648649	0.740957	0.259043	0.351351	0.260824	0.739176	0.739110
DT	Over_0	0.952099	0.729015	0.270985	0.047901	0.159443	0.840557	0.879818
DT	Over_1	0.621622	0.705851	0.294149	0.378378	0.295775	0.704225	0.773462
RF	Train	0.825581	0.726962	0.273038	0.174419	0.271141	0.728859	0.878201
RF	Test	0.810811	0.706915	0.293085	0.189189	0.291080	0.708920	0.834963
RF	Under_0	0.848837	0.767442	0.232558	0.151163	0.191860	0.808140	0.887304
RF	Under_1	0.891892	0.613830	0.386170	0.108108	0.380803	0.619197	0.813894
RF	Over_0	0.928376	0.752965	0.247035	0.071624	0.159329	0.840671	0.889870
RF	Over_1	0.621622	0.735106	0.264894	0.378378	0.267084	0.732916	0.793861
XGB	Train	0.767442	0.691378	0.308622	0.232558	0.307159	0.692841	0.822090

Model	Sampling	TPR	TNR	FPR	FNR	MR	CA	AUC
XGB	Test	0.837838	0.667553	0.332447	0.162162	0.329160	0.670840	0.803601
XGB	Under_0	0.872093	0.627907	0.372093	0.127907	0.250000	0.750000	0.833153
XGB	Under_1	0.918919	0.589894	0.410106	0.081081	0.403756	0.596244	0.798059
XGB	Over_0	0.861314	0.751825	0.248175	0.138686	0.193431	0.806569	0.874696
XGB	Over_1	0.621622	0.734043	0.265957	0.378378	0.268127	0.731873	0.791029
LGBM	Train	0.872093	0.620666	0.379334	0.127907	0.374497	0.625503	0.806339
LGBM	Test	0.918919	0.601596	0.398404	0.081081	0.392280	0.607720	0.809388
LGBM	Under_0	0.883721	0.593023	0.406977	0.116279	0.261628	0.738372	0.822472
LGBM	Under_1	0.918919	0.580319	0.419681	0.081081	0.413146	0.586854	0.786494
LGBM	Over_0	0.861770	0.753422	0.246578	0.138230	0.192404	0.807596	0.883437
LGBM	Over_1	0.621622	0.737234	0.262766	0.378378	0.264997	0.735003	0.783374

Source: Own work