



VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

FACULTY OF BUSINESS MANAGEMENT

DEPARTMENT OF FINANCIAL ENGINEERING

Fabiana Levon

**INVESTIGATION OF FINANCIAL FRAUD DETECTION BY USING BIG
DATA ANALYTICS**

Master Graduation Thesis

Financial Engineering study programme, state code 6211LX060

Financial Technologies (FinTech) specialisation

Finance study field

Vilnius, 2023

<table><tr><td>Vilnius Gediminas Technical University</td></tr><tr><td>Faculty of Business Management</td></tr><tr><td>Department of Financial Engineering</td></tr></table>		Vilnius Gediminas Technical University	Faculty of Business Management	Department of Financial Engineering	<table><tr><td>ISBN</td><td>ISSN</td></tr><tr><td colspan="2">Copies No.</td></tr><tr><td colspan="2">Date-.....-.....</td></tr></table>	ISBN	ISSN	Copies No.		Date-.....-.....	
Vilnius Gediminas Technical University											
Faculty of Business Management											
Department of Financial Engineering											
ISBN	ISSN										
Copies No.											
Date-.....-.....											
Master Degree Studies Financial Engineering study programme Master Graduation Thesis											
Title	Investigation of Financial Fraud Detection by Using Big Data Analytics										
Author	Fabiana Levon										
Academic supervisor	Nijolė Maknickienė										
Thesis language: English											
Annotation											
This thesis focuses on detecting and preventing financial fraud by using big data analytics. The work examines classification techniques as a tool for fraud detection as well as cluster and keyword analysis to distinguish specific characteristics attributed to fraudulent behaviour. Specifically, this study focuses on detecting illegitimate transactions using synthetic dataset of almost 1.9 million US transactions spanning from 2019 until the end of 2020 as well as finding key characteristics associated with fraudulent transactions. The authors find that among three classification techniques – Decision Tree, K Nearest Neighbors and Logistic Regression, the K Nearest Neighbors classifier shows the best performance in terms of detecting fraudulent transactions. Using cluster and keyword analysis it is shown that fraudulent transactions can be characterized most commonly by the transaction categories relating to shopping and groceries. We notice that the percentage of fraudulent transactions committed by women is higher than that committed by men. People performing the fraudulent transactions can be split into two age groups - younger group of ages 33, 40 and 41 and the older group between the ages of 58 and 80. Additionally, by using keyword analysis we find that cities located on the east side of the country are most associated with fraud, especially the city of Washington. Structure: introduction, literature review of financial fraud detection by using big data analytics, theoretical framework and design, study execution and results, conclusions, references. Thesis consist of: 66 p. text without appendixes, 34 illustrations, 5 tables, 63 bibliographical entries. Appendixes of the paper are added separately.											
Keywords: financial fraud, financial fraud detection, big data analytics, classification, cluster analysis, keyword analysis											

Vilniaus Gedimino technikos universitetas
Verslo vadybos fakultetas
Finansų inžinerijos katedra

ISBN	ISSN
Egz. sk.	
Data-.....-.....	

Antrosios pakopos studijų **Finansų inžinerijos** programos magistro baigiamasis darbas

Pavadinimas	Finansinio sukčiavimo atpažinimo tyrimas naudojant didžiųjų duomenų analitiką
Autorius	Fabiana Levon
Vadovas	Nijolė Maknickienė

Kalba: anglų

Anotacija

Šiame darbe pagrindinis dėmesys skiriamas finansinio sukčiavimo aptikimui ir prevencijai naudojant didžiųjų duomenų analitiką. Darbe nagrinėjami klasifikavimo metodai, kaip įrankis sukčiavimui nustatyti bei klasterių ir raktinių žodžių analizė siekiant nustatyti nusikalstamos elgesenos savybes. Šis tyrimas fokusuojasi į neteisėtų mokėjimų aptikimą bei pagrindinių su neteisėtais mokėjimais susijusių savybių paiešką naudojant sintetinius beveik 1.9 milijono JAV mokėjimų duomenis apimančius laikotarpį nuo 2019 m. iki 2020 m. pabaigos. Nustatyta, kad tarp trijų klasifikavimo metodų – sprendimo medžio, K artimiausių kaimynų ir logistinės regresijos, K artimiausių kaimynų klasifikatorius rodo geriausius rezultatus aptinkant neteisėtus mokėjimus. Naudojant klasterių ir raktinių žodžių analizę nustatyta, kad neteisėti mokėjimai dažniausiai naudojami su mokėjimų kategorijomis susijusiomis su bakalėja, apsipirkimu. Pastebėjome, kad moterų nesąžiningų sandorių procentas yra didesnis nei vyrų. Žmonės vykdančius neteisėtus mokėjimus galima suskirstyti į dvi amžiaus grupes – jaunesniąją 33, 40 ir 41 metų grupę ir vyresniąją grupę nuo 58 iki 80 metų. Be to, naudodami raktinių žodžių analizę nustatėme, kad miestai, esantys rytinėje šalies dalyje, yra labiausiai susiję su sukčiavimu, ypač Vašingtono miestas.

Darbą sudaro: įvadas, finansinių sukčiavimo atskleidimo naudojant didžiųjų duomenų analizę literatūros apžvalga, teorinė struktūra ir dizainas, tyrimo vykdymas ir rezultatai, išvados, literatūros sąrašas.

Darbo apimtis: 66 p. teksto be priedų, 34 iliustracijos, 5 lentelės, 63 bibliografiniai šaltiniai.
Atskirai pridedami darbo priedai.

Prasminiai žodžiai: finansinis sukčiavimas, finansinio sukčiavimo aptikimas, didžiųjų duomenų analitika, klasifikavimas, klasterių analizė, raktinių žodžių analizė

TABLE OF CONTENTS

INTRODUCTION	12
1. LITERATURE REVIEW OF FINANCIAL FRAUD DETECTION BY USING BIG DATA ANALYTICS.....	14
1.1. Types of financial fraud	14
1.1.1. Credit card fraud	15
1.1.2. Mortgage fraud.....	15
1.1.3. Money laundering fraud	16
1.1.4. Financial statement fraud.....	17
1.1.5. Securities and commodities fraud.....	18
1.1.6. Insurance fraud.....	18
1.1.7. Cryptocurrency fraud	19
1.2. Methods of detecting financial fraud using Big Data Analytics	21
1.2.1. Mathematical methods.....	21
1.2.2. Artificial intelligence methods.	22
2. THEORETICAL FRAMEWORK AND DESIGN	25
2.1. Classification techniques.....	25
2.1.1. Decision Tree.....	25
2.1.2. Logistic Regression.....	26
2.1.3. K Nearest Neighbors	27
2.2. Assessing the performance of the classification techniques	28
2.2.1. Precision.....	28
2.2.2. Recall	28
2.2.3. F1-Score	29
2.2.4. Accuracy.....	29
2.2. Clustering	30
2.3. Keyword extraction and analysis	31
3. STUDY EXECUTION AND RESULTS	34
3.1. Description of the dataset.....	34
3.2. Application of classifiers.....	38
3.3. Cluster analysis	44
3.4. Keyword analysis	55
CONCLUSIONS	59
REFERENCES	61
APPENDICES	67

LIST OF FIGURES

Figure 1. Methodology scheme	25
Figure 2. K Nearest Neighbors flowchart.....	27
Figure 3. Distribution of non-fraudulent and fraudulent transactions accross the dataset	34
Figure 4. Distribution of transaction time	35
Figure 5. Transaction date of fraudulent and non-fraudulent transactions.....	35
Figure 6. Distribution of transaction amount.....	36
Figure 7. Distribution of fraudulent and non-fraudulent transacions according to gender	37
Figure 8. Distribution of fraudulent and non-fraudulent transactions according to transaction category	37
Figure 9. Distribution of fraudulent and non-fraudulent transactions according to US states	38
Figure 10. Receiver operating characteristic curves (ROC).....	40
Figure 11. Confusion matrix for Logistic Regression	41
Figure 12. Confusion matrix for Decision Tree.....	42
Figure 13. Confusion matrix for K Nearest Neighbors	42
Figure 14. Cluster network	45
Figure 15. Number of observations in first 40 clusters	45
Figure 16. Clusters with the most fraudulent observations	46
Figure 17. Subclusters of cluster 113	46
Figure 18. Transaction date of cluster 113 subclusters.....	47
Figure 19. Customers' birth year across cluster 113 subclusters	47
Figure 20. Transaction categories across cluster 113 subclusters	48
Figure 21. Observations according to gender across cluster 113 subclusters.....	49
Figure 22. Subclusters of cluster 471	49
Figure 23. Transaction date of cluster 471 subclusters.....	50
Figure 24. Customers' birth year across cluster 471 subclusters	50
Figure 25. Transactions categories across cluster 471 subclusters.....	51
Figure 26. Observations according to gender across cluster 471 subclusters.....	51
Figure 27. Subclusters of cluster 270	52
Figure 28. Transaction date of cluster 270 subclusters.....	52
Figure 29. Customers' birth year across cluster 270 subclusters	53
Figure 30. Transactions categories across cluster 270 subclusters.....	53
Figure 31. Observations according to gender across cluster 270 subclusters.....	54

Figure 32. Keyword network visualization|..... 55

Figure 33. Reoccurrence of keywords within fraudulent transactions 56

Figure 34. U.S.A map indicating cities with most association to fraudulent transctions 56

LIST OF TABLES

Table 1. Financial fraud types and related literature with fraud detection methods.....	20
Table 2. Confusion matrix.	28
Table 3. Training scores.....	39
Table 4. Training scores after tuning the parameters.....	39
Table 5. Performance results of different classifiers.	43

LIST OF APPENDICES

Appendix A. Scientific paper published in the research area of the thesis.....	67
---	----

INTRODUCTION

Relevance of the topic.

Financial fraud is a popular topic when discussing various financial institutions, especially the banking and insurance sectors and the newly established Fin Tech sector. According to PwC's (2022) Global Economic Crime and Fraud Survey 2022 "46% of surveyed organisations reported experiencing fraud, corruption or other economic crimes in the last 24 months". To add, financial fraud is extremely costly – companies are losing enormous amounts of money as are their customers. Association of Certified Fraud Examiners (ACFE) in their 2022 report state that organisation lose 5 percent of revenue to fraud each year, causing a total loss of more than 3.6 billion of US dollars(2022). This also leads to distrust between the customer and company, affecting the company's reputation. To prevent such losses the companies must continuously invest in and upgrade their financial fraud battling mechanisms. This is where the usage of Big Data comes in. Big Data analytics is a relatively new field of science which can contribute to detecting fraudulent behaviour. In the age of digitalization, companies providing services online obtain the opportunity to collect massive amounts of data, which in turn can be used and analysed to uncover hidden patterns. Doing this manually introduces additional time, resource and financial costs while introducing Big Data analytics will not only lower the cost, provide more accurate intel, but also improve the overall customer experience. These positive attributes are utilised by companies and businesses dealing with electronic payments and transactions for cost saving and fraud detection benefits. Governments worldwide are also interested in lowering financial fraud rates, therefore information collected regarding fraud is analysed thoroughly to develop fraud prevention strategies and protect people from fraudsters. As stated by Andrew Griffith, Economic Secretary to the British Treasury at the time, regulators and industry experts are working together to ensure that victims of fraudsters are not left bearing the financial burden as payment scams become more sophisticated (Payment System Regulator, 2023). This highlights the importance of cooperation between regulatory institutions and companies – to collect data regarding fraud cases to prevent future fraud as well as educate the public about fraudulent techniques and scams.

As everything around us evolves, including the technologies used to make payments and transactions, so do fraudulent techniques. To detect these new techniques the usage of the modern Big Data tools is needed to inspect and learn of new fraudulent patterns that arise during data analysis and manipulation. Big Data is crucial in this aspect, as other tools simply are not able to process and analyse the enormous amounts of data generated by financial institutions

and businesses – in 2021 over 51 billion credit card payments were performed across the United States as per data reported in Federal Reserve Payments Study (2023). As per the same study the number of transactions keeps increasing each year, therefore usage of Big Data Analytics is essential to analyse and process the collected data.

Problem – Prevention of financial fraud in financial institutions.

Research object – Financial fraud detection techniques and characteristics.

Aim – Examine different techniques for detecting credit card fraud and compare them. Use cluster and keyword analysis to examine characteristics of fraudulent transactions.

Tasks:

1. Describe the theoretical aspects of the chosen fraud detection techniques and tools used to uncover and examine characteristics of fraudulent transactions.
2. Identify types and methods of detecting financial fraud.
3. Apply and compare the selected classification techniques to credit card transactional data and compare their performance of detecting fraudulent transactions. Use cluster and keyword analysis to distinguish characteristics of fraudulent transactions. Discuss usefulness and implementation of the obtained results.

Research Methods. Scientific literature review, comparative analysis of three classification techniques – Logistic Regression, K Nearest Neighbors and Decision Tree, clustering, cluster analysis, keyword analysis.

Authors have also published a scientific article in this field of research pertaining to examination of factors influencing fraudulent transactions using regression and correlation analysis (see appendix A).

1. LITERATURE REVIEW OF FINANCIAL FRAUD DETECTION BY USING BIG DATA ANALYTICS

In this section I aim to identify and briefly describe financial fraud types and briefly and fraud detection techniques that use Big Data Analytics. I will do so by analysing and citing gathered scientific articles.

1.1. Types of financial fraud

Financial fraud is a broad term that refers to financial gain accumulated through the intentional use of illegal methods (West & Bhattacharya, 2016). To better understand what exactly what financial fraud is and how it is executed it can be divided into several types. West & Bhattacharya (2016) firstly defined three common types of financial fraud – bank fraud, corporate fraud and insurance fraud when composing a review of scientific articles on financial fraud detection research from 2004 to 2014. The aforementioned common types of fraud were also distinguished by Al-Hashedi & Magalingam (2021) in their review of data mining techniques applied for financial fraud detection with the addition of cryptocurrency fraud as a new type of fraud.

These common fraud types of financial character are further defined as credit card fraud, mortgage fraud and money laundering fraud – the three of which fall under bank fraud, then there is financial statement fraud as well as securities and commodities fraud, both attributed to corporate fraud and lastly insurance fraud is split into healthcare and automobile insurance fraud (West & Bhattacharya, 2016). The same categories of financial fraud were clarified by Al-Hashedi & Magalingam (2021) with the added seventh type – cryptocurrency fraud, which entails Bitcoin fraud.

Both papers by West & Bhattacharya (2016) and Al-Hashedi & Magalingam (2021) settle on the same categories of financial fraud and both consider insurance fraud as a whole when describing the separate types, combining both healthcare and automobile insurance fraud. Al-Hashedi & Magalingam (2021) also added cryptocurrency as an additional type as well as associate the securities and commodities fraud type with both corporate and insurance fraud. A more in-depth description of the most common types is presented in the following subsections.

1.1.1. Credit card fraud

Credit card fraud can be described as illegal use of a person's credit card for the purpose of conducting fraudulent transactions without the consent of the credit card's owner (West & Bhattacharya, 2016). Al-Hashedi & Magalingam (2021) consider two types of credit card fraud – Online and Offline. In the latter the fraudsters use the physical credit card to conduct illegal transactions without the owner's consent. Online fraud refers to fraud committed remotely, via the Internet or other electronic devices. For detection of Online fraud proactive mechanisms, like high-precision algorithms, are most suitable for real-time detection and prevention, for Offline fraud historical data is used to detect possible missed frauds (Boulieris et al., 2023). Sikkandar et al. (2023) note two types of credit card as well, although different from the previous authors – transaction fraud and application fraud. The latter is associated with identity theft, as the fraudster gets a new credit card using stolen personal information of another person. In the case that an already existing credit card is stolen or a lost card is found with which fraudulent transactions are made – that is considered as transaction fraud.

As Fin Tech sector expands with more payment methods becoming available, the fraudsters are quick to take advantage of them. As electronic banking and other online payment systems gain more and more popularity, the likelihood of fraudulent transactions has increased, resulting in potential losses of billions of dollars annually (Islam et al., 2023). The cost of such type of fraud is beard by both financial institutions issuing the credit cards and the customers. Credit card service providers who do not have a reliable fraud monitoring system are at risk to suffer financial and reputational losses (Hafiz et al., 2016). Islam et al. (2023) raises two concerns which must be addressed by financial organizations – first, default payments due to credit card owners neglecting refunds and secondly, unauthorised third parties performing fraudulent activities.

1.1.2. Mortgage fraud

Mortgage fraud involves individuals targeting mortgage documents and modifying or removing information in these documents during the process of loan application (Al-Hashedi & Magalingam, 2021). By doing so, the value of the property can be misinterpreted, which can influence the lender's decision to fund the loan (West & Bhattacharya, 2016). Nguyen and Pontell (2010) further elaborate on mortgage fraud by unravelling two categories based on the desired outcome – fraud for profit or property. Fraud for property entails less harm, as the applicant conceals certain information to obtain the loan for the property, which he or she intends on repaying. However, the fraudsters who seek profit are in most cases working within

the real-estate industry and have years of experience, which allows them to commit crimes of such nature. Furthermore, this type of fraud has the potential to impact United States' economy as the cost of this type of fraud can be very high. In fact, Baumer et al. (2013) find that foreclosure rates in 2008 were significantly higher in those parts of United States where fraud-for-profit was more common during earlier years, i.e., 2004-2006. In later works Baumer et al. (2017) finds that nearly one fourth of residential mortgage loans from 2003-2005 across America had at least one or more signs of mortgage fraud with varying results across counties. Authors also investigate reasons of why mortgage fraud is committed, as per their findings, loan agents often face situations where they are encouraged to commit fraud either by strict management focused strictly on profits or company culture that promotes such tactics (Nguyen and Pontell, 2010). Ranson et al. (2018) examine factors influencing mortgage fraud in Chicago and find that increased level of mortgage fraud is associated with concentrated disadvantage, concentration of immigrants and black people, though subprime lending mediates the impact of these factors. Authors argue that the relationship between the mortgage fraud and neighbourhood structural characteristics is due to poor public social control in less privileged minority communities. Elul et al. (2023) focused on occupancy fraud, which describes mortgage loan borrowers who misrepresent their occupancy status and found that these types of borrowers perform much worse than similar investors as well as that their defaults were seen as a strategic motive than those of other borrower type. Authors estimate that these types of fraudsters may become a concern in future boom-bust cycles.

1.1.3. Money laundering fraud

Money laundering is a method used by criminals to place money gained through illegal means into valid businesses. It conceals the origin of the money as well as gives off the false appearance of hard-earned and legal income (West & Bhattacharya, 2016). Not only does it make difficult to track down the money and criminals, but it is often used to commit other crimes, such as funding terrorism related actions and weapons trading (Al-Hashedi & Magalingam, 2021). Often there is simply not enough information about a certain organization's transactions to conduct an analysis whether this organization is prone to money laundering (Domashova & Mikhailina, 2021). Antwi et al. (2023) argues that money laundering affects the global financial system by bringing loss of revenue, economic instability and distortions as well as the overall weakening of the financial system as consequences. From a social standpoint, money laundering enables not only drug trafficking and terrorism, but also bribery and corruption, deteriorating the financial system further. To combat the negative consequences, it is necessary that the government steps in to take preventative measures, however some might argue that the

increased transaction costs imposed by anti-money laundering regulations can have negative effects. This claim is debunked by Antwi et al. (2023), stating proof that anti-money laundering policies positively affect the development of financial sector in Africa. Furthermore, a study by Durguti et al. (2023) shows a statistically significant positive effect of anti-money laundering regulations on banking sector stability when examining six Western Balkan countries. To implement effective anti-money policies, Aljassmi et al. (2023) argues that analysis of money laundering and its flows is essential to assess the risk and vulnerability of a specific country to money laundering. Additionally, the analysis of money laundering flows can help to uncover hidden criminal activity generating the funds as well as the actors behind the criminal procedure. Aljassmi et al. (2023) also finds that money laundering is negatively associated with level of financial development as well as gold trade and size of economic activities when performing analysis of factors which influence the impact of money laundering in United Arab Emirates. By determining key factors – real estate market, arms procurement and outflow of money, contributing to the expansion of money laundering in the country, the authors were able to provide concrete suggestions targeting specific channels.

1.1.4. Financial statement fraud

Financial statements contain various information regarding a business' financial activities and conditions – income, profits, loans, ect. They also have certain comments from corporate employees about the business itself that describe the image of the company as well as its financial state (Al-Hashedi & Magalingam, 2021). The main purpose of financial statement fraud is to alter these documents in a way that would make the company seem to be more profitable, have a better status overall (West & Bhattacharya, 2016). This in turn can influence the stock price of the company, reduce tax obligations or simply to gives off the appearance of a highly performing company to please the managers. Kaituko et al. (2023) notes the importance of financial statements especially for publicly listed companies, as the statements are the primary source of information for investors. Worth to note, that this type of fraud is difficult to detect due to it usually being committed by the corporate employees with great understanding and experience in the field of work as well as the infrequency it is committed with (West & Bhattacharya, 2016). Marais et al. (2023) also mentions that out of three major categories – corruption, asset misappropriation and financial statement fraud, the latter is the most costly. Inefficient allocation of resources and financial loss for stakeholders are consequences of poor economic decisions made based on the fraudulent statements. To add, authors differentiate between financial statement fraud and earnings management – although both seek monetary gain, earnings management is conducted within the accounting guidelines,

while financial statement fraud does not comply with accounting standards. Kaituko et al. (2023) show that factors such as a company's board gender diversity, financial expertise, independence and meeting frequency as well as audit fees discourage financial statement fraud in the context of 15 listed manufacturing companies in East Africa. To detect financial fraud, Gepp et al. (2023) argues it is important that the correct variables are considered when modelling this type of fraud, therefore the authors propose changes to the classic fraud triangle framework, which includes pressure, opportunity and rationalisation and present the triangle as combination of pressure/incentive, exploitable opportunity, integrity/attitude/rationalisation along with a new factor – suspicious information, referring to any detectable consequences or data manipulation that occur because of the committed fraud or someone trying to hide the fraud.

1.1.5. Securities and commodities fraud

Definition for this type of fraud varies significantly across different academic fields, Yu (2013) uses a rather loose definition of the term investing corporate security fraud, which implies material loss experienced by stakeholders due to misconduct from the company or its management. This particular type of fraud uses false information to deceive a person into investing money into a company by various methods. This includes Ponzi Schemes, Pyramid Schemes, Embezzlements and others (West & Bhattacharya, 2016). Authors mention, that this particular type of fraud has not been covered by a lot of studies, from those that did take interest in this type of fraud, most inspect external forms of it, while inspecting the cases of its internal form. Interesting study by Cumming et. al (2015) was conducted on corporate securities fraud and gender diversity, where the authors find that the proportion of women on boards helps to reduce the frequency and severity of fraud cases, attributing this to the higher level of risk aversion in women. However, Steffensmeier et al. (2015) found that higher percentage of women was arrested due to occupational crime, which includes embezzlement, than men.

1.1.6. Insurance fraud

Insurance fraud aims to gain illegal profit from the insurance company by committing a fake accident, injury or loss of assets. It misuses insurance policies that are intended to protect businesses from financial losses and takes advantage of them (Al-Hashedi & Magalingam, 2021). There are different types of insurance fraud – in relation to health insurance fraudsters can deceive the insurance company by placing a bill for expensive medical procedures. In automobile fraud fabricated documents referring to a staged car accident can be submitted (Al-Hashedi & Magalingam, 2021). Crop insurance fraud overstates the losses suffered by

customers due to natural disasters or diminishing prices (West & Bhattacharya, 2016). Warren & Schweitzer (2018) study consumer insurance fraud and find that interpersonal avoidance is a key determinant of unethical behaviour and that interviews with the claimant are a must in order to determine the legitimacy of a claim. Fiederling et al. (2018) tested the consumer behaviour towards insurance fraud using a questionnaire and an experiment, where the authors found that the answers provided in the survey do not match the behaviour displayed in the experiment.

1.1.7. Cryptocurrency fraud

As stated by Al-Hashedi & Magalingam (2021) cryptocurrency is often used by criminals due to lack of regulation and decentralization. This type of fraud deceives people with promises of big profits while providing them with fake investments or services. It mainly targets people with inadequate knowledge about the market and takes advantage of their naivety. These kinds of fraud can generate revenues worth millions of dollars (Al-Hashedi & Magalingam, 2021). Patsakis et al. (2023) argues that intensive regulation by governments will not help to mitigate the risks associated with cryptocurrencies effectively. Additionally, the authors find proof that fraudulent transactions via cryptocurrency can be traced back to the fraudsters by examining illicit bitcoin payments carried out by ransomware groups. The paper shows that a large amount of bitcoin transactions can be traced back to the fraudsters quite easily, thus enabling the tracking and investigation of these criminals and providing opportunities for developing legislation, which will be better and fairer to cryptocurrency. Authors also stress that in order to pertain this traceability it is important not to push strict regulation, as this might incentivise the fraudsters to move their operations to a different, better-secured domain, where the traceability might be harder. Cong et al. (2022) consider Wash-trading, when investors sell and buy the same asset at the same time, thus affecting the volume, price and volatility as well as investors' confidence and involvement in financial markets. The study found that signs of wash trading on most unregulated exchanges, where the volumes reported were inflated by over 70 percent, on average. The increased volume is shown to be utilized for temporary price distortion, exchange ranking improvement (Cong et al., 2022). Kerr et al. (2023) explores the risks of cryptocurrencies by analysing major fraud scandals, highlighting the FTX scandal of 2022, as reasons that governments are considering placing regulatory restrictions and oversight mechanisms in the industry. Additionally, the authors stress that although the prices of cryptocurrencies are similar to that of company stocks in that both can rise and fall unexpectedly, the risk are much higher in the case of cryptocurrency as its value is essentially zero, unlike the company stock value which depends on the profits of a company. Nghiem et

al. (2021) examine a similar type of cryptocurrency fraud – pump-and-dump schemes, which involves inflating the price of the cryptocurrency and selling it immediately to unsuspecting traders. The procedure is initialised with the help of social media, where on platforms such as Telegram or Discord, the communication regarding the planned time and cryptocurrency is taking place. As such fraudulent tactics and information about cryptocurrencies is freely available on the Internet, young and impressionable investors are at risk - Ryan (2023) argues that The Securities and Exchange Commission should regulate exchanges by placing an age limit and having customer identification systems in place to make sure that minors are not able to purchase cryptocurrency without consent of an adult. Author recalls incidents where social media influencers, who have many children watching them, promote the unregulated digital currency to their audience, then sell large amounts of the currency for profit, which lowers the value of the coin significantly.

Table 1. Financial fraud types and related literature with fraud detection methods

Type of financial fraud	Literature	Methods
Credit card fraud	Saia and Carta (2019), Madhurya et al. (2022), Shabbir et al. (2022), Chen et al. (2015), Dong et al. (2021) and others.	Fourier transform and Wavelet transform, Support Vector Machine, quantum neural network, RAIN model, ect.
Money laundering	Singh and Best (2019), Fronzetti and Remondi (2017)	Visualization techniques, network analytic techniques
Financial statement fraud	Zhang et al. (2022), Cheng et al. (2021), Shen et al. (2021), Jan (2021)	Support Vector Machine, Random Forest, Naive Bayes, Decision tree, Logistic regression, Recurrent Neural Network, Long short-term memory
Insurance fraud	Bologa et al. (2013), Yan et al. (2020), Amponsah et al. (2022)	Social network analysis, Kernal Ridge Regression, decision tree

Source: designed by author

Table 1 presents a brief overview of scientific literature and fraud detection methods used categorized according to types of financial fraud. The fraud detection techniques will be presented in more detail in the following subsection.

1.2. Methods of detecting financial fraud using Big Data Analytics

As financial fraud is not relatively new, various questions and methods regarding its prevention and detection have been raised and analysed. However, this study is focusing on the use of Big Data – as West & Bhattacharya (2016) have put it, Big Data is less expensive and time consuming as well as more accurate than traditional methods like manual detection. The techniques used for fraud detection can be split into two groups – mathematical and artificial intelligence based. The two categories and different techniques are described below.

1.2.1. Mathematical methods.

The logistic model is used to examine the association between free and discrete ward factors (Bi & Liang, 2022). When used to detect financial statement fraud the logistic regression method had the lowest accuracy of 59 percent out of the five approaches considered (SVM, KNN, DT, NB and LR) at the baseline. When knowledge graph models were considered, the accuracy for the logistic regression improved, however, not enough to become the best performing model in terms of accuracy (Shen et al., 2021). LR method together with text analytics was used by Dong et al. (2018) to detect corporate fraud. This model was also used to detect loan fraud, where its accuracy was only around 87 percent, one of the lowest of the five models (NB, NN, RF, SVM and LR) evaluated in the study (Attigeri et al., 2021). Li (2022) tested the LR model when dealing with e-commerce fraud, the accuracy resulting between 70 and 90 percent, depending on the sample size.

Saia and Carta (2019) evaluated two new strategies that focus on proactive fraud detection. The two strategies considered by the authors are Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) processes. The first strategy, Discrete Fourier Transform (DFT), focuses on defining the model in terms of frequency components, while the other is used to move the evaluation process to a new time-frequency-domain (Saia and Carta, 2019). The authors used transactions from two datasets – European Transactions and German Credit, to test the two previously mentioned approaches and compare them Random Forests approach. The results showed that the Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) strategies were able to perform as well as Random Forests approach. These two approaches deal with two problems that are faced during fraud detection using other methods – data scarcity and unbalanced datasets, since they do not contain previous fraudulent transactions (Saia and Carta, 2019). Authors also suggest implementing the aforementioned strategies in a hybrid environment with other approaches to utilize the advantages of the DFT and DWT.

1.2.2. Artificial intelligence methods.

Decision tree (DT) is a graphical strategy, where features are classified at each node. One of the most important features when building a decision tree is determining the branching criteria (Bi & Liang, 2022). It is useful when solving complex problems (Cheng et al., 2021). In the paper investigating financial statement fraud by Shen et al. (2021), one of the five classifiers considered was Decision tree, which had the highest accuracy percentage when looking at the baseline results. Nonetheless, when knowledge graph models were introduced, its accuracy dropped from nearly 68 percent to around 63 percent (Shen et al., 2021). Amponsah et al. (2022) investigated fraudulent behaviour in the healthcare industry using decision tree algorithm and blockchain technology. Different type of decision trees were considered, pruned and unpruned. Results stated that the model in question accurately classified health insurance claims 98 percent of the time (Amponsah et al., 2022). The DT approach was also utilized in combination with text analytic framework by Dong et al. (2018) to detect corporate fraud.

Support Vector Machine (SVM) are a supervised type of Machine Learning technique. By using a certain kernel function they map data to another high dimensional space and is able to find the hyperplane with the biggest margin between two classes (Madhurya et al., 2022). When testing this approach for classifying fraudulent credit card transactions it was proven to be efficient, although possible problems arising from unbalanced and differing datasets were raised (Madhurya et al., 2022). Attigeri et al. (2021) found that SVM approach to detecting loan fraud based on Non-Performing Assets was not the most effective as its accuracy was the lowest out of all the models considered. It was also shown to have the lowest results when performance for imbalanced datasets and specificity was considered (Attigeri et al., 2021). Jain et al. (2022) also found that SVM approach for identifying credit card fraud did not yield the highest accuracy rate. Dong et al. (2021) researched fraud detection in the e-market using framework which was based on SVM. It was concluded that SVM based model decreased the error rate for around 20 percent when comparing to other approaches (Dong et al., 2021). Shen et al. (2021) compared SVM method to other approaches when evaluating models for financial statement fraud. Comparatively, SVM method was not the best performing in terms of accuracy when baseline results were considered, however, with addition of knowledge graph models it had the highest results for accuracy percentage, F-measure and AUC. Dong et al. (2018) also used SVM approach in combination with text analytic framework to detect corporate fraud. A similar study with the use of SVM and text analytics for detecting fraudulent statements was conducted by Zhang et al. (2022) yielding in a best performing model in terms of accuracy (71

percent). Li (2022) utilized SVM when examining e-commerce fraud and found that depending on the sample size, the accuracy of the model was between 60 and 90 percent.

Random forest (RF) is a learning algorithm. At its core it is a Bayes classifier, an implementation of the previously mentioned Decision trees (Madhurya et al., 2022). It is a collection of many Decision trees, where each decision tree has a random factor (Cheng et al., 2021). When used to detect credit card fraud, the accuracy of the model was the highest out of six approaches considered – around 80 percent. Because of the imbalanced datasets, pre-processing and sampling algorithms are needed to classify the raw data before applying the Random Forest technique (Madhurya et al., 2022). Attigeri et al. (2021) considered Random forest when looking at loan fraud. This strategy was used to predict the outcomes as „Performing Asset“ or „Non-performing Asset“. Its accuracy was calculated to be 91 percent and an F1 Score as well as Precision indexes over 0.9 (Attigeri et al., 2021). Such method is less likely to suffer from overfitting and noise effects (Cheng et al., 2021). Zhang et al. (2022) combined text analytics and Random Forest approach to test for fraudulent statements, resulting with an accuracy rate of 66 percent.

The Neural Network (NN) model with many layers made up of artificial neurons (Attigeri et al., 2021). This approach had the highest accuracy (around 94 percent) among other models (NB, LR, RF, SVM and NN) in detecting loan fraud. Its precision and specificity indexes were the highest as well (Attigeri et al., 2021). Dong et al. (2018) applied the Neural Network approach together with text analytics to test for corporate fraud.

Recurrent Neural Network (RNN) is great for analysing the relationships between data points (Jan, 2021). Considering financial statement fraud, the RNN model was accurate 87 percent of the time and its precision was almost 87 percent. Long short-term memory model (LSTM) is an extension of the aforementioned RNN model that performs better than the original RNN in terms of long-term memory (Jan, 2021). Indeed, as the results show the LSTM model has a higher accuracy (95 percent) and precision (92 percent) rate.

Naive Bayes (NB) is a predictive model that uses prior and likelihood probabilities (Attigeri et al., 2021). When dealing with loan fraud, its accuracy was 92 percent and its precision index was above 0.9. Comparatively it was the second best model out of five (LR, NN, RF, SVM and NB) with regards to not only accuracy, but precision as well (Attigeri et al., 2021). This approach was used to detect credit card fraud as well, where its accuracy was much lower (76 percent) when compared to other models considered in the same study (SVM. Fusion model and NB) (Jain et al., 2022). Naive Bayes was also used along text analytics by Zhang et al. (2022) to detect fraudulent statements with quite low accuracy (56 percent) when compared to other models that were mentioned (Bag of Words, RF, SVM and NB).

Social Network Analysis is based on retrieving additional data on the relationships of a subject. It creates a link between different data sources and allows for forecasting fraudulent behaviour as opposed to just detecting it (Madhuri et al., 2021). Firstly, the relationships between major system information entities are modeled as a network, then suspicious components are identified via set of indicators and later the suspects can be detected by performing simulations (Bologa et al., 2013). This method takes into consideration the fact that usually fraud is performed by a group (network) of people, not just a single individual. So, by detecting one suspicious person, the group (network) of this person can be analysed to filter other frauds (Bologa et al., 2013).

Singh and Best (2019) tackled the problem of detecting money laundering by using visualization techniques to detect unusual bank transactions. Yan et al. (2020) focused on the application of Kernel Ridge Regression, specifically Artificial Bee Colony algorithm based, for detecting automobile insurance fraud – the proposed model with the addition of Artificial Bee Colony algorithm reports an accuracy percentage above 97 percent. Chen et al. (2015) studied fraudulent transactions using RAIN (Risk of Activity, Identity and Network) model to quantify the risk of an object or user. Shabbir et al. (2022) examined suspicious transactions using quantum neural network with achieved accuracy of 97 percent. Leite et al. (2020) proposed NEVA (Network Detection with Visual Analytics) to detect patterns of fraudulent behaviour in bank transactions. Fronzetti and Remondi (2017) used network analytic techniques (social network metrics) to detect money laundering activities. To spot clusters of potential criminals visual techniques were utilized as well.

To sum up, main types of financial fraud were identified and main methods for detecting financial fraud were analyzed with the help of scientific literature analysis. Examining Table 1 we can conclude that not enough research is conducted on the following types of financial fraud – mortgage, securities and commodities as well as cryptocurrency fraud. Furthermore, not all methods of fraud detection are analyzed in each fraud category. For example, although Support Vector Machine approach is popular, in the literature reviewed, it is only used to detect credit card and financial statement fraud. The scientific articles used for analysis mainly use accuracy percentage to evaluate the models. Other indexes used for evaluating the approaches include specificity, precision and F score. A problem concerning the availability of data for research purposes was also raised in the literature – it is hard to get data containing sensitive financial information because of security and privacy reasons.

2. THEORETICAL FRAMEWORK AND DESIGN

In this section I will present and analyse the methodology that will be used to execute the empirical study. The first subsection goes over the chosen classification techniques that will be analysed throughout the work and the following subsection will be regarding performance criteria. The third subsection introduces the clustering procedure and its framework, while the next subsection dives deep into keyword analysis and its details. Detail scheme of the investigation is provided by Figure 1.

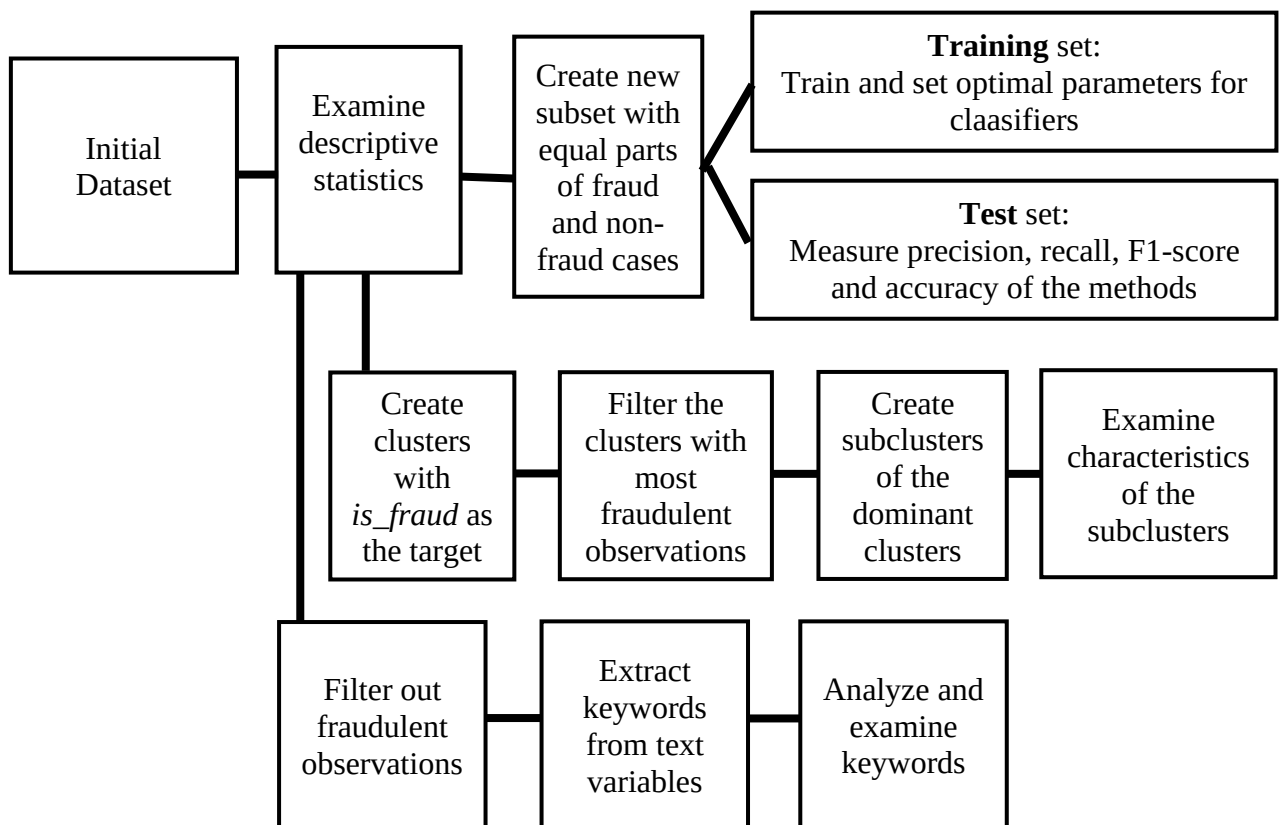


Figure 1. Methodology scheme

Source: designed by author

2.1. Classification techniques

2.1.1. *Decision Tree*

This classifier is often used in scientific papers regarding fraud detection and is praised for its simplicity in implementation and understanding (West & Bhattacharya, 2016). Although it is regarded as simple, the technique is powerful as it has been applied to solve complex

problems of artificial intelligence. It is a prediction model illustrated as a hierarchical tree structure, which does not require any domain knowledge or parameter setting for its construction, therefore, this technique is knowledge discovery (Cheng et. al, 2021). The decision tree is composed of nodes, which encapsulate specific linear test functions. The functions are defined by characteristics of the dataset – time, amount of the transaction for example. By constructing sequences of rules, the decision tree can adapt the training sample by locating a critical pattern (Chen et al., 2018). However, both West & Bhattacharya (2016) and Chen et al. (2018) mention the issue of overfitting when working with this classifier when faced with an extremely complex problem and large datasets.

2.1.2. Logistic Regression

Logistic Regression – statistical method of classifying binary data by executing a regression on a set of variables. Combining the input variables containings certain characteristics with a response variable, it utalizes the natural logarithm and calculates probability that the result is within the bounds of a certain category. For binary classification, as is the case in this study, the response variable takes on the following form (West & Bhattacharya, 2016):

$$y_i = \begin{cases} 1 \\ 0 \end{cases}, \quad (1)$$

here:

y_i - denotes whether a transaction i has been recognized as fraudulent or not.

Finally, the probability of sample x_i belonging in class 1 is calculated by equation 2 (West & Bhattacharya, 2016):

$$P(Y_i = 1|X_i) = \frac{\exp(\omega_0 + \omega^T x_i)}{1 + \exp(\omega_0 + \omega^T x_i)}, \quad (2)$$

here:

ω_0 - the parameter representing the intercept;

ω^T - coefficient vector.

Although similar to Decision Tree classifier the Logistic Regression is simple implement, West & Bhattacharya (2016) also mention that this particular classification technique has poor classification performance when compared to other methods.

2.1.3. *K Nearest Neighbors*

The particular classification technique is not often referred to in the literature that has been analyzed, however, in the few works (Shen et al., 2021; Madhurya et al., 2022) where it has been utilized, K Nearest Neighbors has shown impressive performance results.

The classifier works according to the following logic – all of the instances used in the training procedure are plotted and the labelled instance is used to classify the unlabelled instance based on the nearest neighbor (Madhurya et al., 2022). A more detailed view of how the K Nearest Neighbors classifier works is depicted in Figure 2:

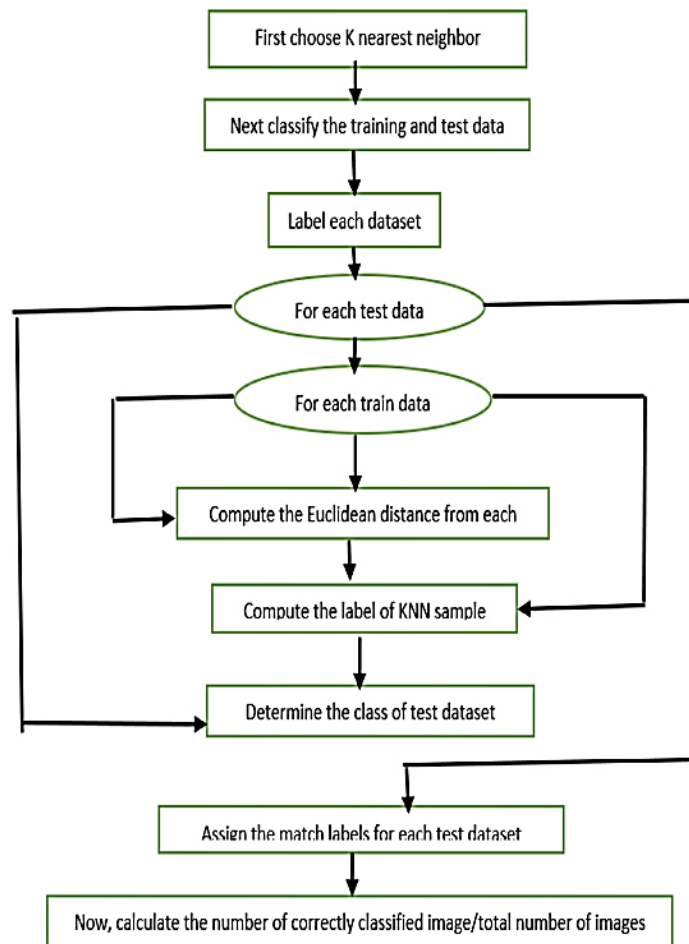


Figure 2. K Nearest Neighbors flowchart

Source: Madhurya et al. (2022)

These three classification techniques were chosen deliberately – both the Decision Tree and K Nearest Neighbors classifiers were chosen because of their documented high performance in detecting fraud, one method – the Decision Tree, is often used in scientific works and the K Nearest Neighbors is not as popular among authors. Finally, Logistic Regression is also analysed, this provides an opportunity to validate the findings of other

authors regarding the comparatively low performance of this technique and see how its performance compares to the two aforementioned methods.

2.2. Assessing the performance of the classification techniques

To understand how the classification techniques perform in detecting fraudulent transactions the following metrics are calculated and examined. The metrics require understanding of concepts such as true negative, false positive, false negative and true positive from the confusion matrix, which are explained by table 2:

Table 2. Confusion matrix

	Predicted value 0 (non-fraud)	Predicted value 1 (fraud)
True value 0 (non-fraud)	True Negative	False Positive
True value 1 (fraud)	False Negative	True Positive

Source: Zhang et al. (2022)

2.2.1. Precision

This metric refers to the proportion of correctly judged fraud samples out of all samples predicted as fraudulent and calculated by the following equation (Zhang et al., 2022):

$$Precision = \frac{True\ positives}{True\ positives + False\ Positives}, \quad (3)$$

This measure is not the most commonly used measure by authors in their scientific works, although Zhang et al. (2022), Domashova & Mikhailina (2021) as well as Attigeri et al. (2021) have utilized this measure along with others to compose a full picture of the performance of different methods.

2.2.2. Recall

Recall, or sometimes referred to as Sensitivity (Amponsah et al., 2022), is the true positive rate, the proportion of correctly labelled transactions as fraudulent to the total amount of transactions that are fraudulent within the set, given by (Zhang et al., 2022):

$$Recall = \frac{True\ positives}{True\ positives + False\ Negatives}, \quad (4)$$

This metric indicated how well the technique is able to capture the fraudulent cases. In comparison, Recall is more commonly used than Precision by scientists, such as in the papers of Amponsah et al. (2022), Dong et al. (2018), Domashova & Mikhailina (2021) and others.

2.2.3. *F1-Score*

Combining the previously mentioned precision and recall measures, the F1 score measures the trade-off between the two measures, calculating the harmonic mean of the two according to equation 5 (Zhang et al., 2022):

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

F1-Score is frequently used to assess a classifiers or models performance alongside accuracy and recall measures, as did Dong et al. (2018) and Attigeri et al. (2021) in their works.

2.2.4. *Accuracy*

This parameter evaluates the overall accuracy of the proposed method by accounting for the proportion of correct predictions for both groups to the total number of predictions (Madhurya et al., 2022):

$$Accuracy = \frac{True\ positives + True\ Negatives}{True\ positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (6)$$

This measure is by far the most popular when reading scientific articles. Most of the articles reviewed for this study have used this metric for model evaluation purposes as it gives an overall accuracy rating for a method, rather than metrics such as recall, which only take into account the correctness of labelling fraudulent transactions. However, as Madhurya et al. (2022) has mentioned, this parameter can be misleading, especially when credit card transactions are involved, as the amount of illegitimate transactions is usually a lot lower than the entire amount of transactions, otherwise referred to as an imbalanced dataset.

2.2. Clustering

To study attributes of the dataset relating to fraud, clustering mechanism is utilised in this study. In its essence, clustering aims to divide objects or observations into conceptually meaningful clusters or groups. The grouping procedure seeks that the objects or observations in one group are similar to one another, while separate clusters are different from each other (Ngai et. al, 2011).

To execute the clustering, Uniform Manifold Approximation and Projection (UMAP) graph-based technique is used to improve outlier detection and accuracy performance. This is a method used to reduce dimensionality – taking datapoints from high-dimensional space to low-dimensional space, while preserving the quality of the presentation. The reduction in dimensionality is useful for easier outlier detection, especially in datasets with high-dimensional features or multiple variables, such as will be used in this work (Setiawan et. al, 2023). The UMAP method is based on three assumptions (Bhowmik et. al, 2022):

1. Even distribution of the datapoints over a Riemannian manifold
2. Riemannian metric is locally constant
3. Riemannian manifold is locally linked.

The UMAP algorithm itself is composed of two parts – first, fuzzy topological representation must be constructed. In practice, the fuzzy set membership strengths have to be calculated for a point's nearest neighbour, because they decrease to be vanishingly small. To swiftly compute the nearest neighbours in high dimensional spaces, Dong et al. (2011) Nearest-Neighbour-Descent algorithm is utilised, so that the remaining computing is done solely with the established local neighbour of each data point (How UMAP Works — Umap 0.5 Documentation, n.d.).

Second part is the optimization of the low dimensional representation so that it is as similar to the fuzzy topological representation as possible, given the cross entropy:

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right) \quad (7)$$

Here:

$w_h(e)$ - constitutes the weight functions of the 1-simplex, e , for the high dimensional cases;

$w_l(e)$ - constitutes the weight functions of the 1-simplex, e , for the low dimensional cases;

E – denotes the set of all possible 1-simplices.

This equation should be minimized, which can be described in two parts – the first, minimizing the first component of the equation when the $w_l(e)$ term is the largest, while the second component will be minimized when $w_l(e)$ is as small as possible. This creates a pull effect from each side, so that the representation with the reduced dimensionality settles in such a state, which accurately represents the overall topology of the initial data.

In adapting this optimization mechanism for real-life usage, some shortcuts were made. As the function needs to be differentiable, an approximation of the actual membership strength function for the low dimensional representation is used – UMAP utilizes curves of the form $\frac{1}{1+ax^{2b}}$. Additionally, spectral embedding techniques are used to set the low dimensional representation into a suitable state (How UMAP Works — Umap 0.5 Documentation, n.d.).

To create the clusters with the transactional data Graphext (<https://www.graphext.com/>) software is used, to group the data using machine learning technology. Target and factor variables have to be chosen to compute the clusters – based on the factor variables the similarity is determined. Similarity is calculated between rows of data and use this information to split the observations into clusters. As we are utilising the UMAP method, our target variable indicating fraud will also influence the clusters (Graphext, Models, n.d.). The factor variables are chosen according to correlation coefficient, that is, variables with the highest correlation with the variable of interest, *is_fraud*, are chosen to be the factor variables.

After the clustering procedure is executed, the network is visualized via graph. The clusters are later analysed, by selecting the clusters with most fraudulent transactions we can examine these clusters further by creating subclusters and analysing different characteristics of the illegitimate transactions themselves and people performing the transactions.

2.3. Keyword extraction and analysis

In addition to clustering, keyword analysis can be used to further characterize fraudulent transactions, especially when having a large dataset with text variables, as is the case with this paper.

To extract the keywords from text variables, Graphext software is utilized – the Natural Language Processing (NLP) algorithms allow the user to extract language features, which in turn can be analysed further. NLP is a branch of computer science, which focuses on understanding and processing human language while being able to showcase the understanding in terms of statistical models using machine learning. As computers cannot read the text

variables the same way as humans, transforming the data so that it becomes readable for the computer is important. There are quite a few ways that text transformation is executed, for e.g. count vectorization, Bag-of-words and others, which represent the text data in the form of numerical values or vectors (Graphext, An Introduction to Natural Language Processing for Text Analysis, n.d.). Graphext defines keywords as significant terms contained within the analysed text, which are later clustered in terms of how related the keywords are. Therefore, first, the software extracts the keywords from each row of data and based on how often the word is repeated either keeps it or neglects the keyword. Then, the keywords are transformed into vectors, which represent them as numbers – this is used to link words with respect to semantic similarity (Graphext, Keywords, n.d.).

Within the model specification in the software, one must choose whether the keyword analysis shall focus on speed or precision, as we aim to analyse the keywords which best describe the fraudulent transactions, we prioritize precision. Other options (Network configuration and visualization) are not altered and the procedure is executed with the default settings within the software, therefore the keywords which are repeated less than three times are dismissed. After the procedure three variables are created – count, keyword and cluster. The first variable is used to apply size mapping to nodes when creating the network visualization, while the cluster variable groups related keywords using the Louvain algorithm (Graphext, Keywords, n.d.). The Louvain algorithm is based on hierarchical clustering aiming to maximise the modularity of community partition by going through continuous iterations of mobile nodes and retrieve the optimal community partition (Zhang et. al, 2021). The main metric of this algorithm, modularity, evaluates the quality of community partition, given by the following formula (Zhang et. al, 2021):

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j) \quad (8)$$

here:

k_i - the degree of node i ;

C_i - shows the community of node i ;

A_{ij} - represents the weight of the edges connected between nodes i and j ;

m - is the weighted sum of all edges in the network;

$\delta(C_i, C_j)$ - takes on the value 1 if the the nodes i and j are in the same community and 0 otherwise.

In conclusion, the chosen classifiers were presented – decision tree, logistic regression and k nearest neighbors have been presented from a theoretical framework. The performance of the chosen classification techniques will be judged based on the metrics described in the subsection two – precision, recall, F-1 score and accuracy calculated according to the presented formulas. Additionally, confusion matrixes are created and their results are taken into account. Secondly, clustering is discussed – using Graphext software and utilizing UMAP dimensionality reduction technique clusters and subclusters shall be created and examined to detect characteristics of fraudulent transactions. Finally, keyword analysis execution framework is presented – utilizing natural language processing algorithms, keywords are extracted and clustered using the Louvain algorithm. Mentioned keywords are analysed based on repeated frequency and used to uncover characteristics of illegitimate transactions.

3. STUDY EXECUTION AND RESULTS

This chapter presents the investigative part of the research and its results – firstly, the data and its descriptive statistics are analysed, further the classification techniques are applied and their performance is examined. Lastly, cluster and keyword analysis is performed to analyse the traits of fraudulent transactions.

3.1. Description of the dataset

In order to create a well-performing fraud detection model, it would be ideal to have real-life transactional data from a financial institution or company. This would allow to study and capture any specific behavioural patterns that may be present within the transactional data for a specific company. However, as mentioned in the second section of this paper, financial data of this character is not widely available, therefore, this study makes use of a publicly available dataset of simulated credit card transactions generated by Sparkov Data Generation found on the Kaggle (<https://www.kaggle.com/>) platform.

The initial dataset consists of 1 852 394 transactions performed across United States of America performed from the 1st of January, 2019 until December 31st, 2020. The transactions are described by 23 variables, out of which 11 are numeric – consecutive transaction number, credit card number, transaction amount, unix time, variables describing the location such as zip code, latitude and longitude, city population, merchant's latitude and longitude and the binary variable indicating whether a particular transaction is fraudulent. The other 12 variables are categorical describing the specific time of the transaction as well as its unique number containing letters and other characteristics such as the name and category of the merchant, full name, gender, state, city, street, occupation and date of birth of the person performing the transaction. Among these variables there are no missing values.

Firstly, I examine the amount of fraudulent and non-fraudulent transactions.

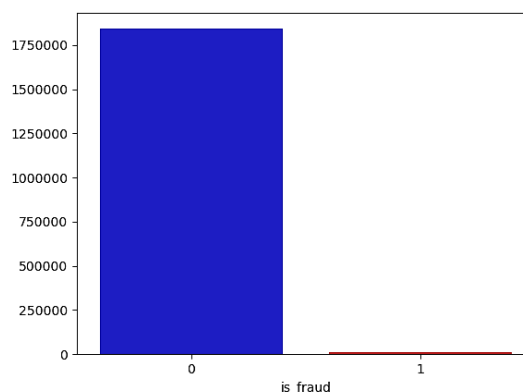


Figure 3. Distribution of non-fraudulent and fraudulent transactions

Source: designed by author

As is visible from Figure 3, the legitimate transactions (presented by the blue column) outweigh the number of fraudulent ones, in fact, only 0,52 percent of total transactions are fraudulent. This constitutes an imbalanced dataset problem.

Next, I analyze the distribution of the variables indicating the transaction amount and time. For the analysis of the time aspect, I will be using the unix time¹, as it is given in a numerical format in the initial dataset and easier to work with from a computational standpoint.

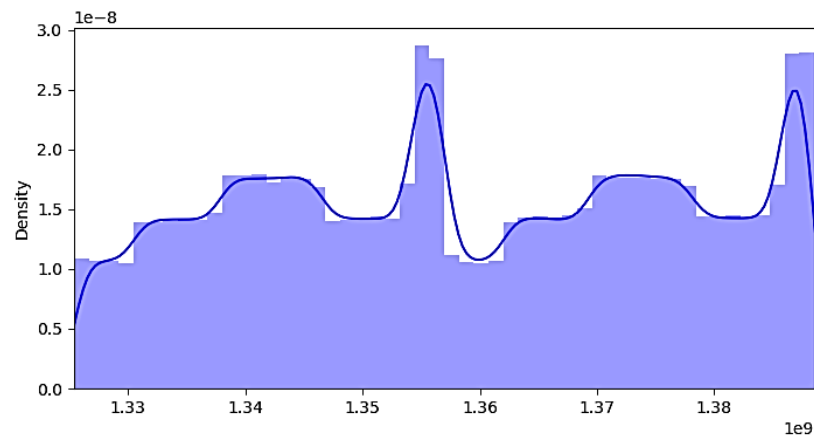


Figure 4. Distribution of transaction time

Source: designed by author

Inspecting Figure 4, the distribution of transactions time is not normal. Two peaks of transactions can be seen of the figure where up to 3 transactions were performed at the exact same time, while in other cases the number of simultaneous transactions taking place is lower.

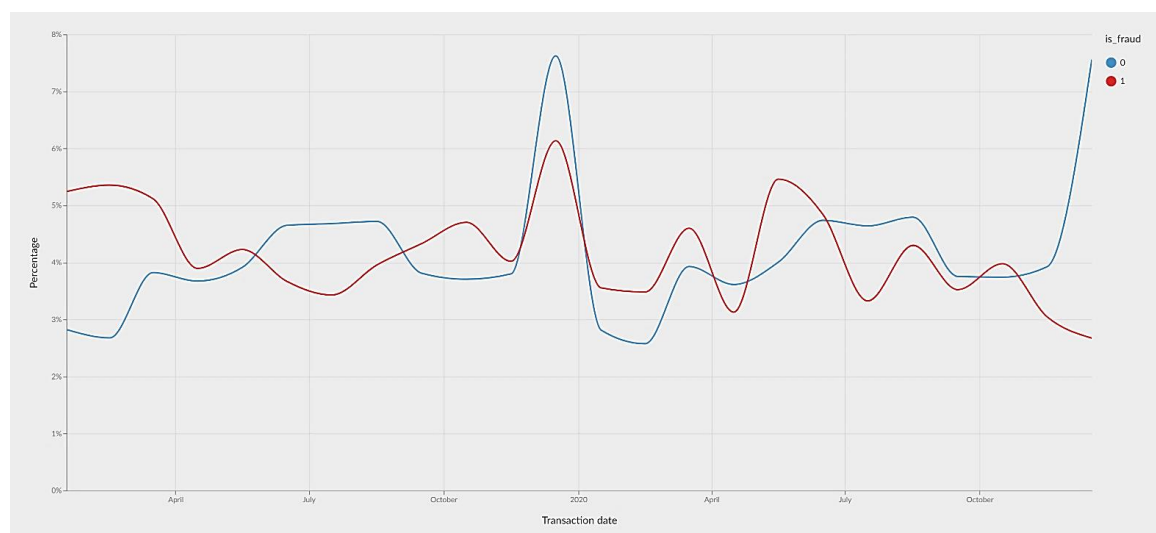


Figure 5. Transaction date of fraudulent and non-fraudulent transactions

Source: designed by author

¹ Unix time – method of describing a point in time. The count starts at the Unix Epoch on 1970-01-01 at UTC and counts how many seconds have passed from this starting point until a certain event, in this case the transaction occurrence (UnixTime.org)

The above figure shows the percentage of fraudulent and non-fraudulent transactions across dates. During 2019 the trends for fraudulent and non-fraudulent go in opposite directions – the percentage of fraudulent transactions went up in the first half year of 2019 and the end of 2019 while outweighing the proportion of legitimate transactions. During the increased number of transactions during year end of 2019 and 2020 – the proportion of non-fraudulent transactions is higher, although in 2019 the percentage of fraudulent transactions also went up, but in 2020 it showed a downturn trend.

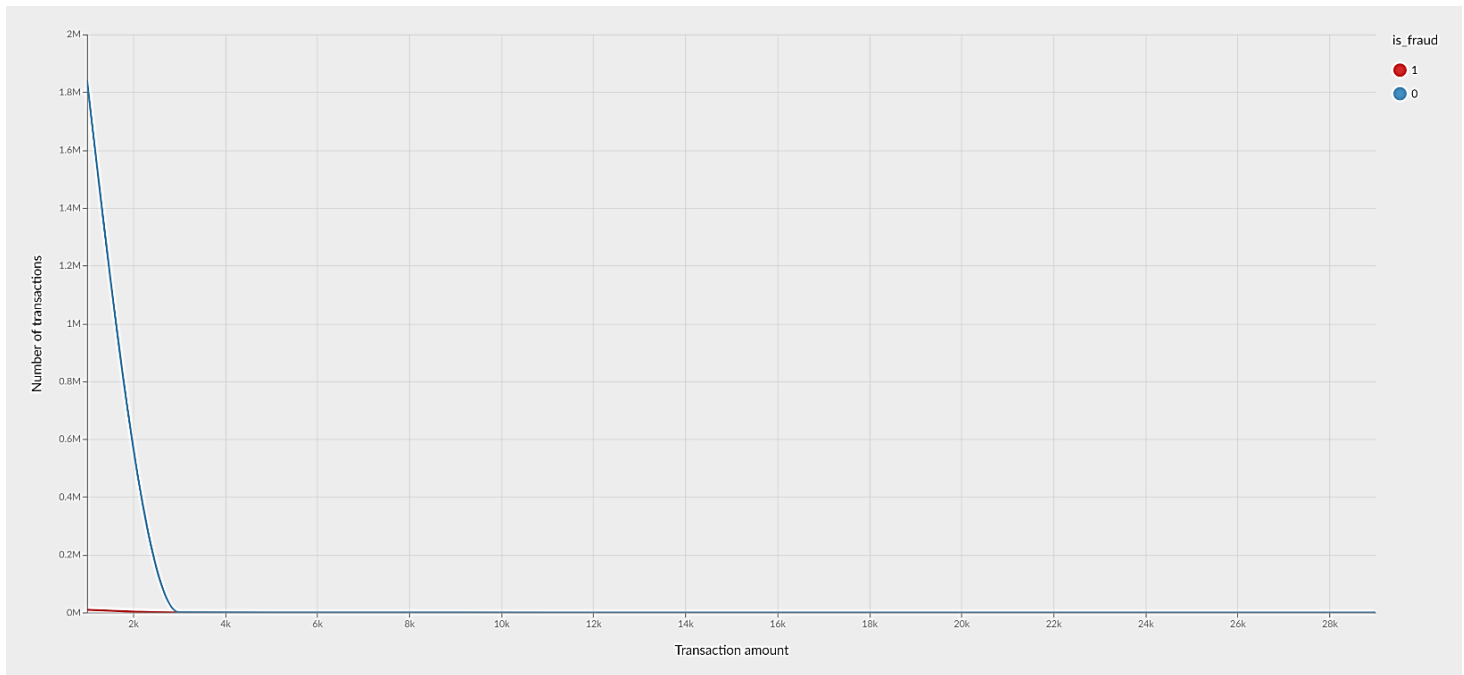


Figure 6. Distribution of transaction amount

Source: designed by author

Figure 6 examines the distribution of transaction amount. The range for the amount is quite large as the smallest transaction amount is only 1,00 and the largest is 28 948,90, however we see that the distribution is skewed to the left, indicating that the majority of transactions performed are closer to the minimum value – from 0 to 3 000. This is again confirmed as the average amount of the transaction is only 70,06 and the median is equal to 47,45. As majority of the transactions are legitimate, their range is much wider. The fraudulent transactions have a much more narrow range – with the smallest being 1,06 and the largest – 1 376, 04.

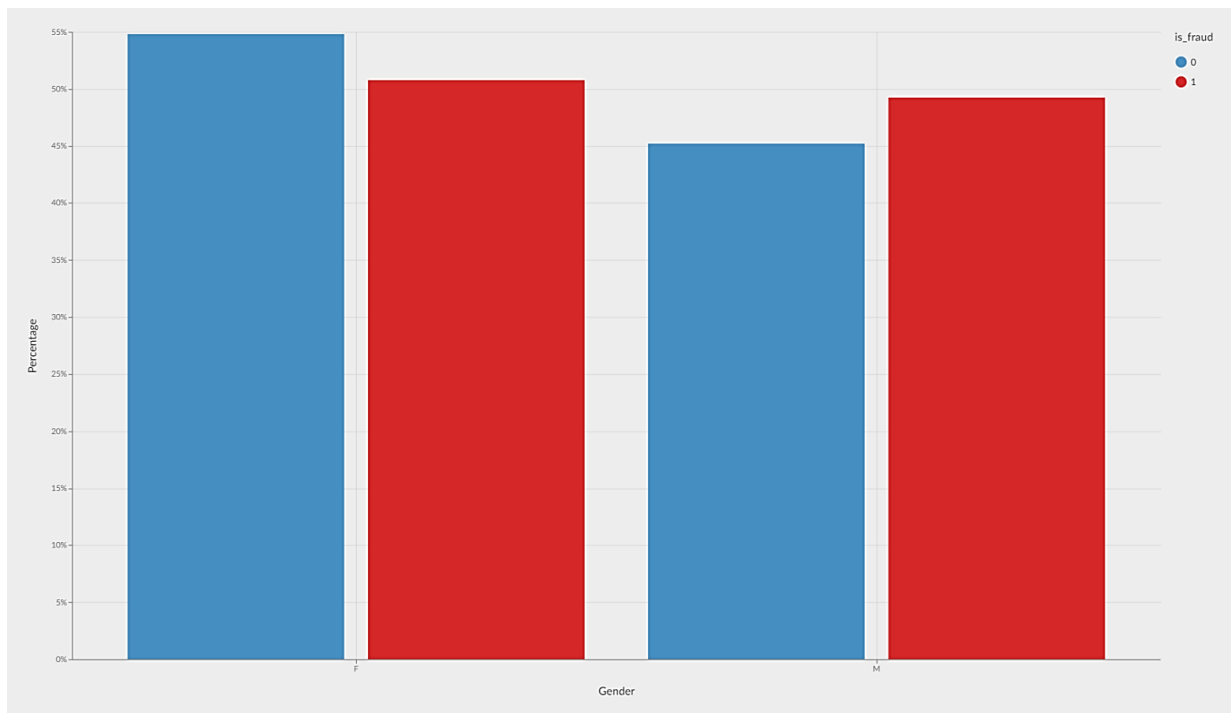


Figure 7. Distribution of fraudulent and non-fraudulent transactions according to gender

Source: designed by author

Figure 7 shows the percentage of fraudulent and non-fraudulent transactions for men and women. Number of transactions for both categories are higher for women – as women are responsible for 54,78 percent of all transactions. Note, the portion of fraudulent transactions outweighs the portion of legitimate transactions for men – 49,24 percent of all transactions made by men are fraudulent.

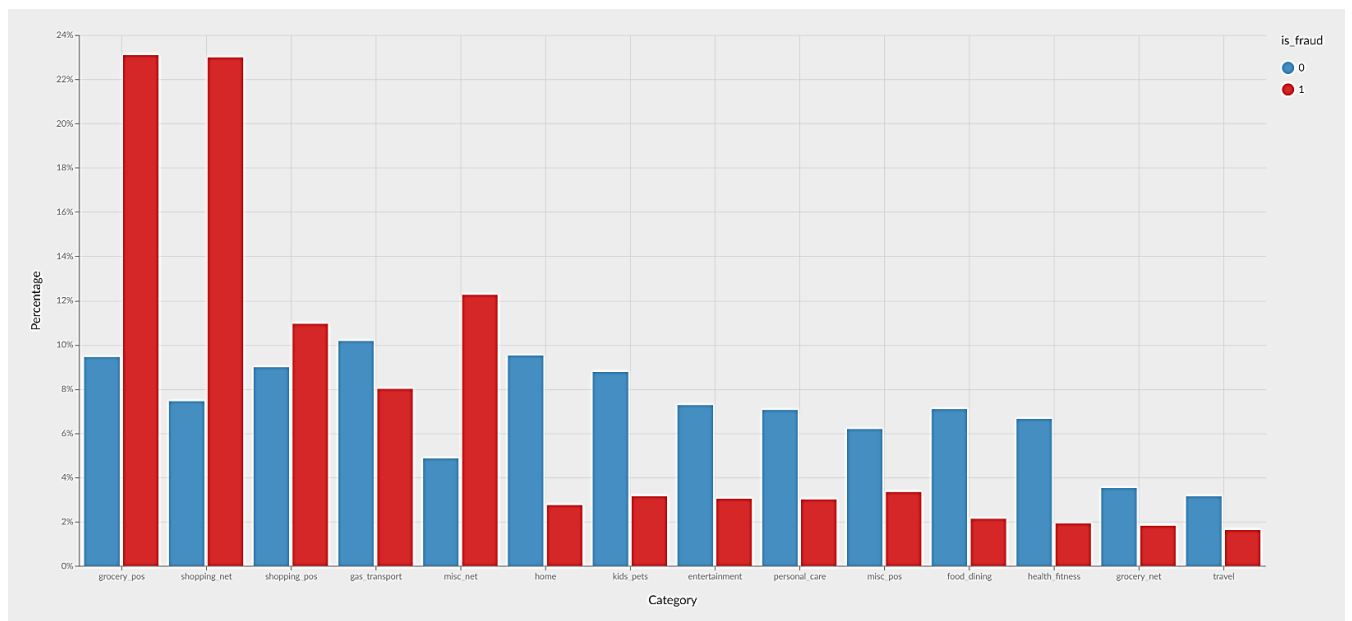


Figure 8. Distribution of fraudulent and non-fraudulent transactions according to transaction category

Source: designed by author

Figure 8 presents the distribution of fraudulent and non-fraudulent transactions according to transaction category. The biggest concentration of illegitimate transactions in ‘grocery pos’ category, which stands for ‘point of sale’ transactions made in the grocery category. The second largest portion of illegitimate transactions is attributed to the ‘shopping net’ category. Important to notice that in categories ‘shopping pos’ and ‘misc net’ also have larger shares of fraudulent transactions – 10,94 and 12,25 percent, respectfully.

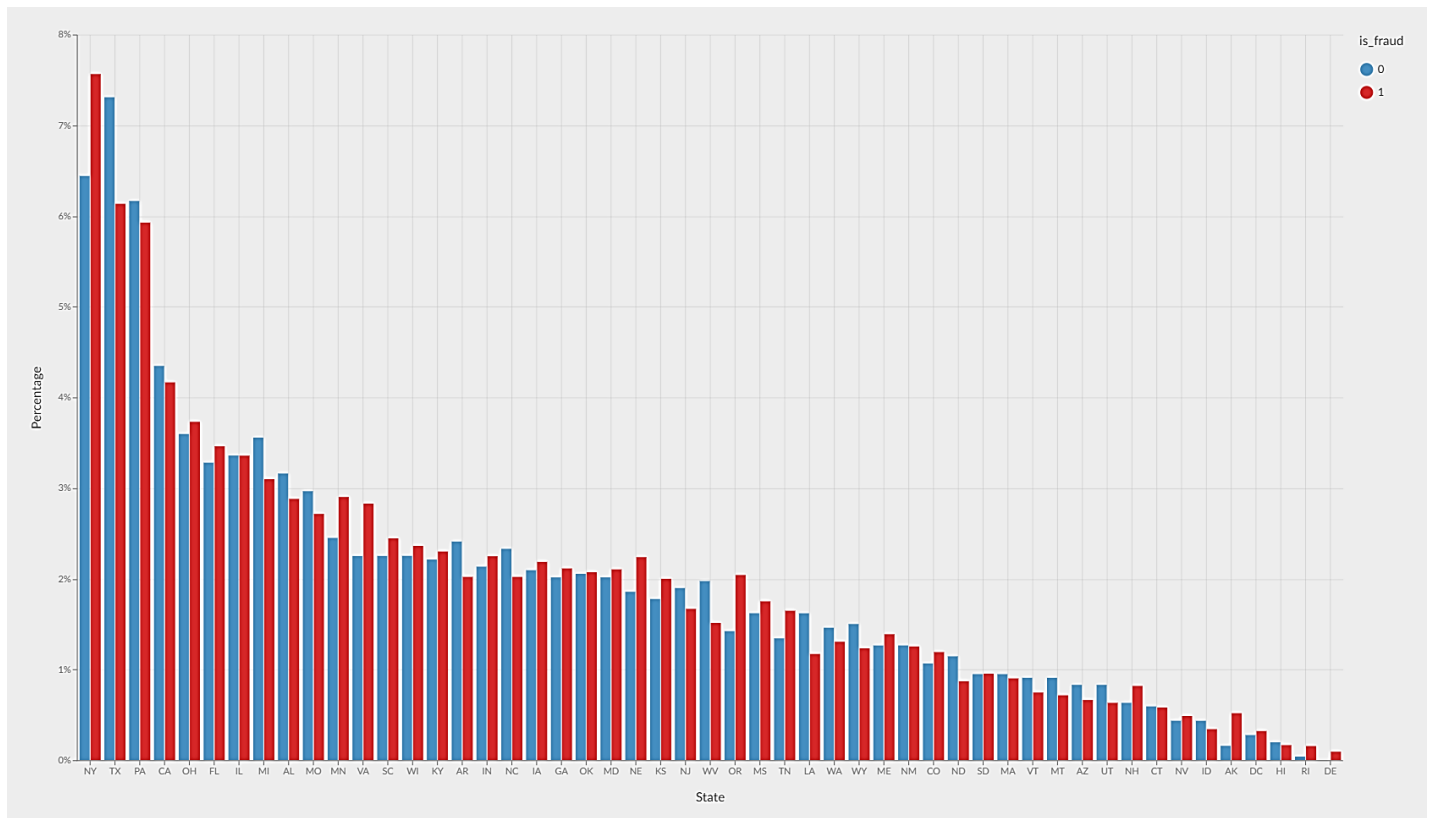


Figure 9. Distribution of fraudulent and non-fraudulent transactions according to US states

Source: designed by author

Figure 9 shows the distribution of fraudulent and non-fraudulent transactions according to US states. Biggest proportion of fraudulent transactions was made in the state of New York. Texas and Pennsylvania have the second and third largest share of fraudulent transactions. Most of the legitimate transactions were made in Texas, second and third are the states of New York and Pennsylvania.

3.2. Application of classifiers

As the distribution of fraudulent class is skewed, a new dataset is generated with equal distribution of both types of transactions – all of the 9 651 illegitimate transactions are taken and combined with the same amount of randomly selected non-fraudulent transactions. The

transactions are shuffled and the new dataset containing equal parts of fraudulent and non-fraudulent is created. Additional variables of scaled transaction amount and scaled unix time are added to standardize the dataset. Outliers for transaction amount and time are removed as well. For the execution of the classifier analysis, only numerical variables are left, the categorical characteristics are erased for computational purposes.

After the removal of outliers, the new dataset containing 19 297 transactions is split into training and test sets – 15 437 transactions are reserved for training and the remaining 20 percent is the testing set, the same split ratio used by Zhang et al.(2022). After, the classifiers are applied to the training samples. Training score is calculated using the accuracy metric:

Table 3. Training scores

Classifier	Accuracy, %
Logistic Regression	50,0
Decision Tree	86,0
K Nearest Neighbors	79,0

Source: compiled by author.

As depicted in Table 3, during initial training Logistic Regression had the lowest accuracy value, while Decision Tree was the most accurate in assigning the correct labels to the transactions.

Next, I try to optimize the parameters of each classifier to increase the predictive power of the methods by using GridSearchCV library function from Python.

Table 4. Training scores after tuning the parameters

Classifier	Accuracy, %
Logistic Regression	50,22
Decision Tree	87,51
K Nearest Neighbors	92,56

Source: compiled by author

The results from Table 4 suggest that parameter tuning helped to better the performance of K Nearest Neighbors classifier by a significant margin, while the increase for the Decision Tree classifier was a lot less, the Logistic Regression results improved by less than a fourth of a percent.

To further analyse the performance of the classifiers with the training set, the Receiver Operating Curve (ROC) is plotted and based on it the Area Under the Curve (AUC) is calculated. The false positive rate on the X axis of the ROC curve represents for the portion of wrongly classified fraudulent transactions as non-fraudulent in the entire non-fraudulent category. The positive rate on the Y axis indicates the proportion of fraudulent transactions predicted as fraudulent out of all actually fraudulent transactions. If the curve for a particular classifier is higher than the diagonal line – it is better at classifying than a random classifier. The higher the curve the better the results. AUC measure shows the exact area under the ROC curve, so that many classifiers with similar results can be compared (Zhang et al., 2022).

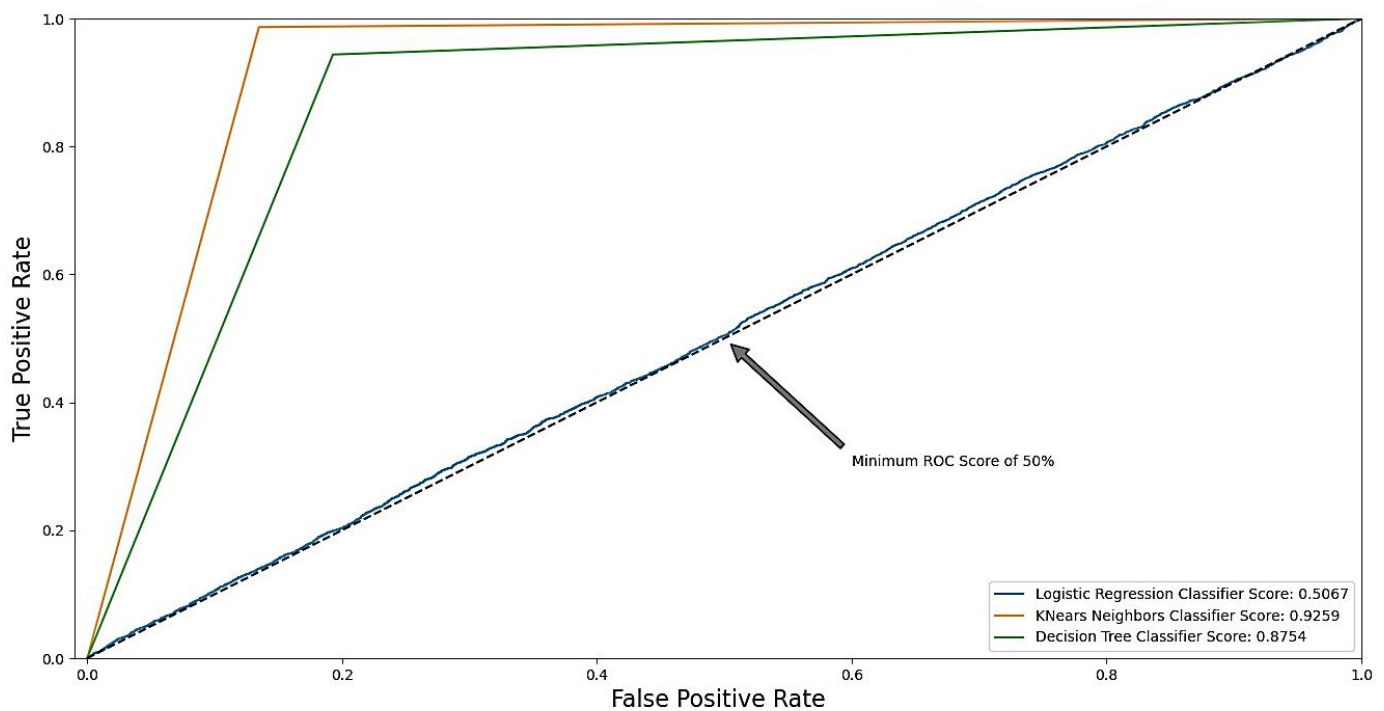


Figure 10. Receiver operating characteristic curves (ROC)

Source: designed by author

Figure 10 clearly indicates the two best performing classifiers – K Nearest Neighbors and Decision Tree. The ROC curves for both are way above the diagonal line and subsequently their AUC scores are very high – 0,93 and 0,88 respectively. Logistic regression scored poorly in terms of AUC (0,51) as its ROC curve is almost following the diagonal line in the graph. These results are comparable to the training scores given in table 4, both in performance rating and numerical results.

Now, that the training is done, we can apply the classification techniques on the test data and evaluate the results.

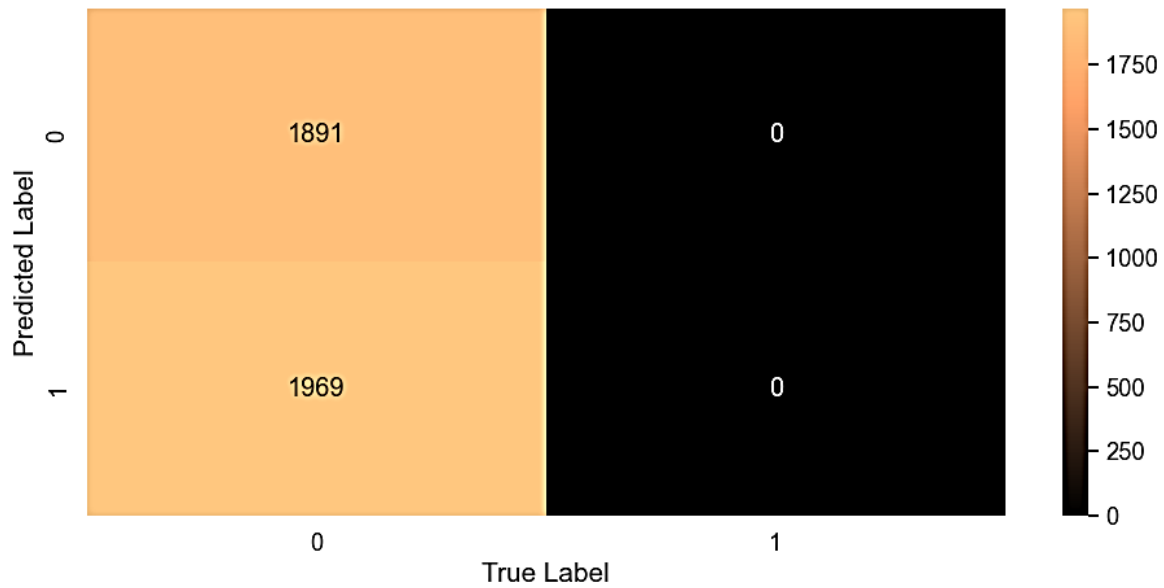


Figure 11. Confusion matrix for Logistic Regression

Source: designed by author

Figure 11 presents the confusion matrix for Logistic Regression. This classifier was not able to assign the fraudulent label to a single transaction correctly. As the point of a fraud detection mechanism is to correctly detect the fraudulent transactions, based on the given results, the Logistic Regression classifier would perform terribly. However, this method did correctly assign the non-fraudulent label to a large number of transactions(1891). 1 969 transactions that were legitimate were classified as fraudulent, this instance describes the type I error accounting for all the false positives. Using the equations for types I and II errors provided by Cheng et. al. (2021), the type I error is estimated at 51 percent. As there were no false negatives, the type II error is equal to 0. This is important as reiterated by the aforementioned authors, type II error has a much higher misclassification cost than type I error. False negatives refer to transactions that have been marked by the detection mechanism as non-fraudulent even though they are fraudulent. This implies that in real life scenario, these transactions would be fully processed without any screening and could bring major financial and reputational losses for the company or business.

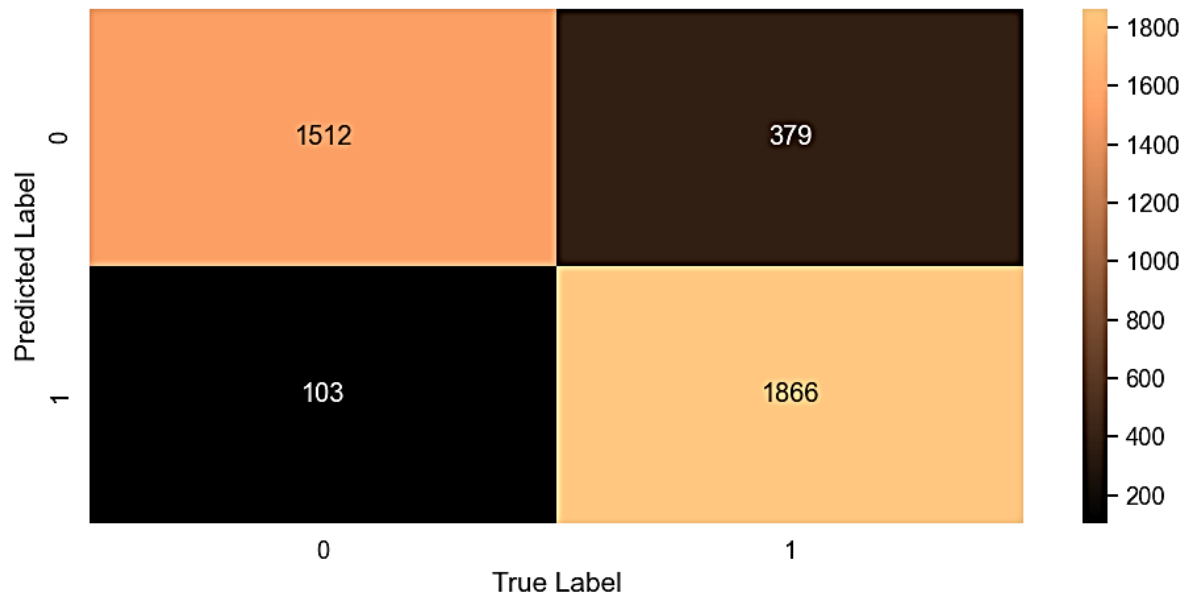


Figure 12. Confusion matrix for Decision Tree

Source: designed by author

Next, I examine the confusion matrix for the Decision Tree classifier given in Figure 12. We see that a large amount(1866) of fraudulent transactions was caught by the classifier. This method also proved useful to correctly label the majority of legitimate transactions. The false positives, or type I error is calculated to be equal to roughly 2,7 percent. In turn, the type II error affects around 9,8 percent. Constituting that almost a tenth of transactions were misclassified with high probability of legal and financial repercussions.

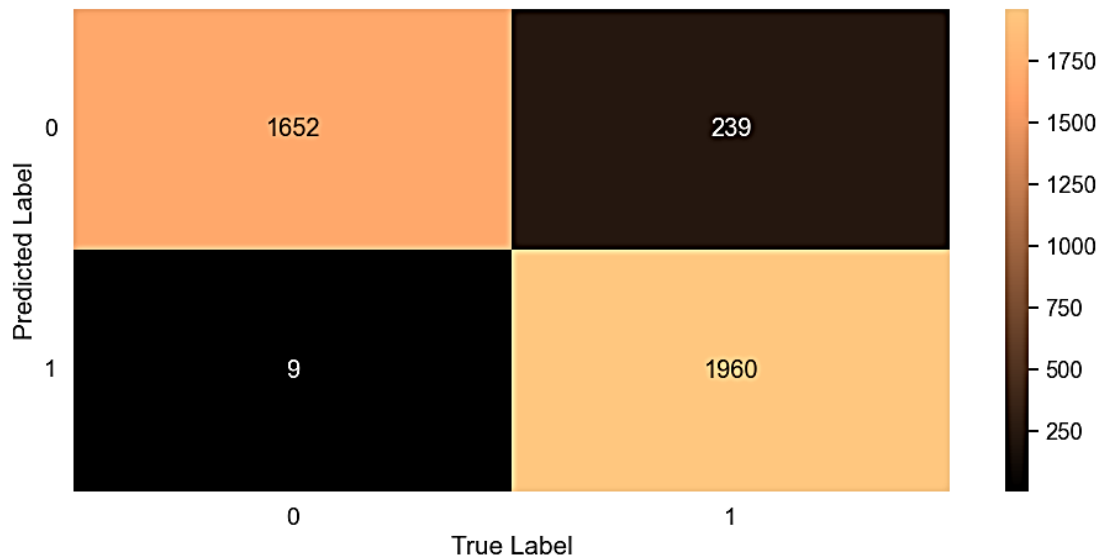


Figure 13. Confusion matrix for K Nearest Neighbors

Source: designed by author

According to the confusion matrix for the K Nearest Neighbors (Figure 13), this method

is well adapted to find and correctly classify the majority of legitimate and illegitimate transactions. Type I error in this instance equals to only 0,23 percent, while type II error is estimated at 6,2 percent. Low error rates and high number of correctly labelled transactions speak in favour of this method as best suitable for fraud detection.

Finally, I analyse the Table 5 containing our main metrics for assessing the performance of these classifiers.

Table 5. Performance results of different classifiers

	Logistic Regression			Decision Tree			K Nearest Neighbors			Support
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	
0	0,49	1,00	0,66	0,94	0,80	0,86	0,99	0,87	0,93	1891
1	0	0	0	0,83	0,95	0,89	0,89	1,00	0,94	1969
Accuracy			0,49			0,88			0,94	3860
Macro avg	0,24	0,50	0,33	0,88	0,87	0,87	0,94	0,93	0,94	3860
Weighted avg	0,24	0,49	0,32	0,88	0,88	0,87	0,94	0,94	0,94	3860

Source: designed by author

According to all of the selected performance metrics the K Nearest Neighbors classification technique is the best performing method in terms of detecting fraudulent transactions. It was able to correctly identify almost 90 percent of transactions as fraudulent while capturing all of the illegitimate transactions that were in the dataset. Therefore, it also was a high F1-score (0,94) regarding the fraud detection portion. Decision Tree also displayed high performance metrics while detecting fraudulent transactions – it was able to detect 95 percent of 1969 total fraudulent transactions with precision of 83 percent, concluding with a relatively F1-Score as well (0,89). As mentioned before, the Logistic Regression did not perform as well as the other two methods in detecting and correctly classifying fraudulent transactions as it was not able to classify a single fraudulent transaction correctly.

If we focus on the correct labelling for the legitimate transactions, which there are 1891 of in the test data, the K Nearest Neighbors classification technique still performs the best – capturing 87 percent of all the legitimate transactions out of which 99 percent were correctly labelled as non-fraud cases. Decision Tree also shows great results in classification of legitimate transactions with 80 percent captured during the classification process with high precision (94 percent). Logistic Regression was able to correctly classify all of the non-fraudulent

transactions, however, because it had no luck in identifying the fraudulent transactions, its precision rate dropped to half.

Table 5 also presents macro averaged scores which equal weight for each class, regardless of the ratio dictated by the supporting observations. The weighted average results are also included. All of the previously mentioned conclusions and results still stand true for both the macro averaged and weighted scores.

After looking at individual classifiers results' in terms of detecting separately fraudulent and non-fraudulent transactions we can finish by reviewing the overall accuracy. No surprise that K Nearest Neighbors has the highest accuracy score (94 percent) as it was mentioned to be great in classifying both classes of transactions, Decision Tree has the second highest accuracy score (88 percent), while Logistic Regression is by far the least accurate out of the three.

In comparison, Shen et al. (2021) also performed analysis of five classifiers, among which the three used in this study are present. In their work, the results are defined by the accuracy percentage as well as F1-score with baseline and feature selection models considered. In both scenarios, all of the three techniques' scores are comparable – the scores for Logistic Regression are not significantly lower as this study has found. The accuracy rating fluctuates between 60 and 68 percent – higher than the results for Logistic Regression in this study, yet lower when accuracy of the Decision Tree and K Nearest Neighbors are taken into account. The F measure in Shen et al. (2021) paper is between 0,56 to 0,7, whereas in this case study, it is again much higher if Logistic Regression is put aside. Although the results for the three classifiers are quite similar for Shen et al. (2021), K Nearest Neighbor classifier has marginally better accuracy and F1-score when feature embedding results are taken into account. Important to note that Shen et al. (2021) was researching financial statement fraud not credit card fraud, the difference in the type of fraud analysed can be attributed to the difference in results.

3.3. Cluster analysis

To perform cluster analysis, firstly, I identify five variables with the highest correlation to the variable of interest *is_fraud*. In this dataset these were the variables indicating the transaction category, unix time, transactions date and time, customer's date of birth and merchant. These variables were added as factors, while the target variable is the binary variable indicating fraud, to form clusters.



Figure 14. Cluster network

Source: designed by author

Figure 14 presents the clusters formed, in all – 1 763 clusters. The first ten, according to the number of observations, are highlighted.

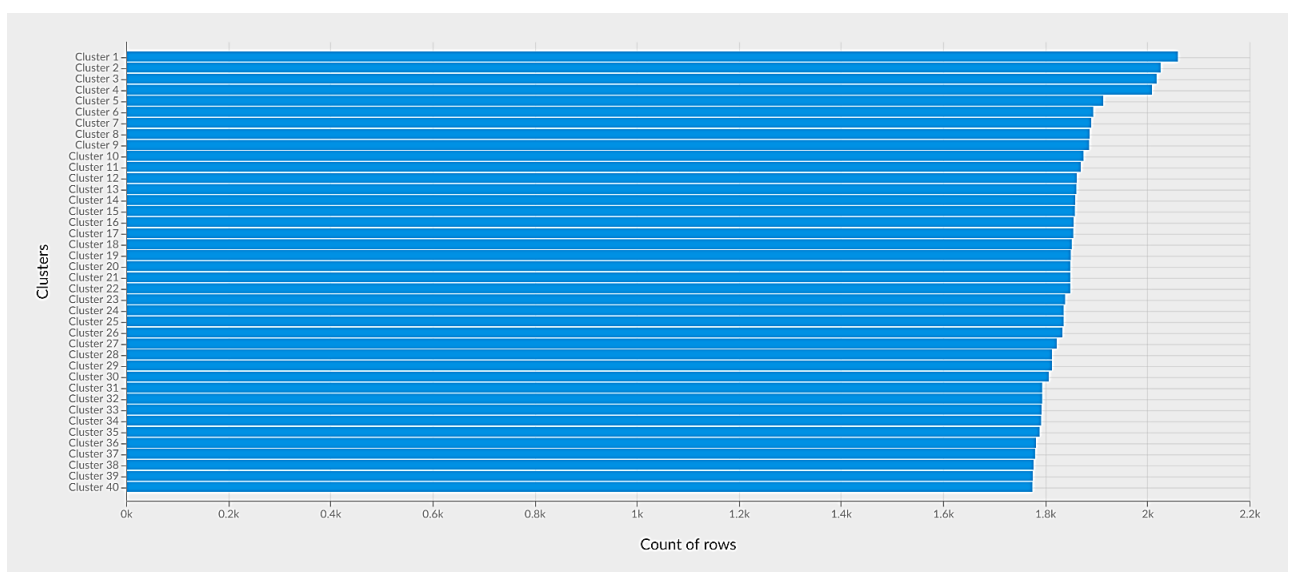


Figure 15. Number of observations in first 40 clusters

Source: designed by author

Figure 15 shows the first 40 clusters and number of observations in each. The first four clusters are special in terms of number of observations – all of them have over 2 000 transactions, while the following clusters face a sharp decrease in number of observations.

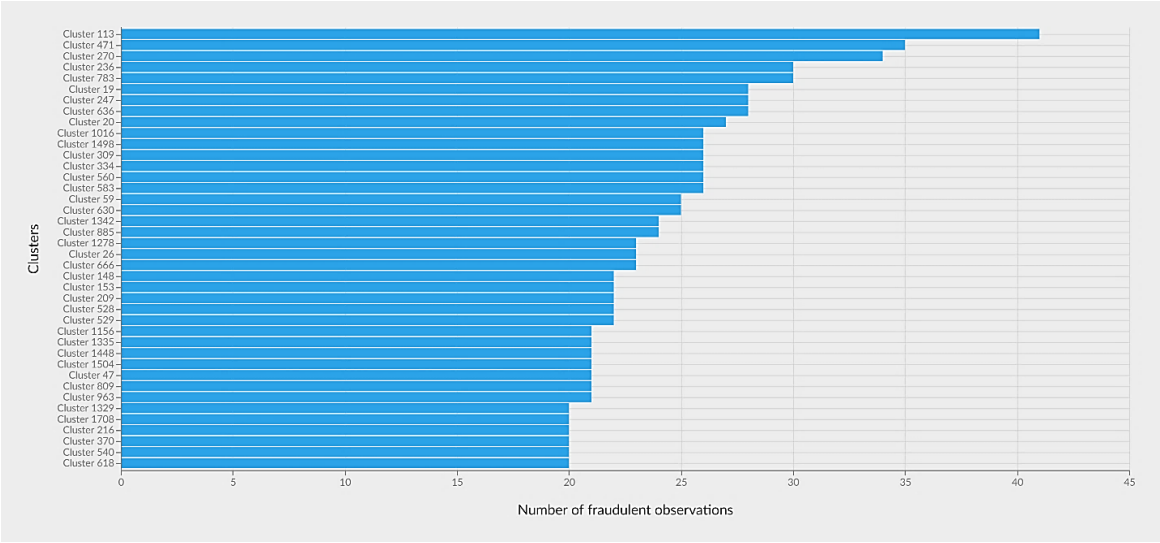


Figure 16. Clusters with the most fraudulent observations

Source: designed by author

The above graph presents the order of clusters according to the number of fraudulent observations. Cluster 113 has 1 593 observations, out of which 41 are fraudulent, 471st cluster has 1 287 observations, 35 being fraudulent, cluster 270 - 1 435 transactions and 34 are illegitimate, clusters 236 and 783 have 1 465 and 1 102 observations, respectively, and both include 30 fraudulent ones. All other clusters have less than 30 fraudulent transactions observed. To further analyse the clusters, I take the first three clusters with most fraudulent observations and compile subclusters to examine characteristics of the illegitimate transactions.

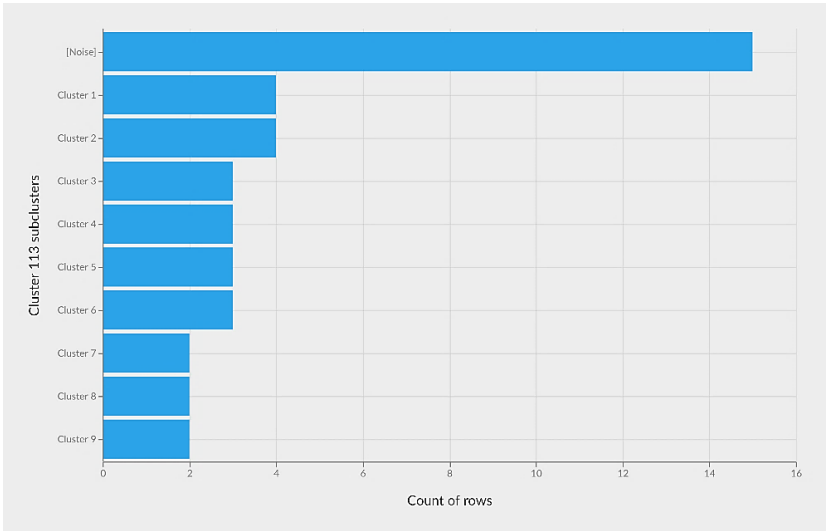


Figure 17. Subclusters of cluster 113

Source: designed by author

Figure 17 presents the subclusters of cluster 113 – all 41 fraudulent observations have been divided into 9 subclusters with a special subcluster with noisy datapoints.

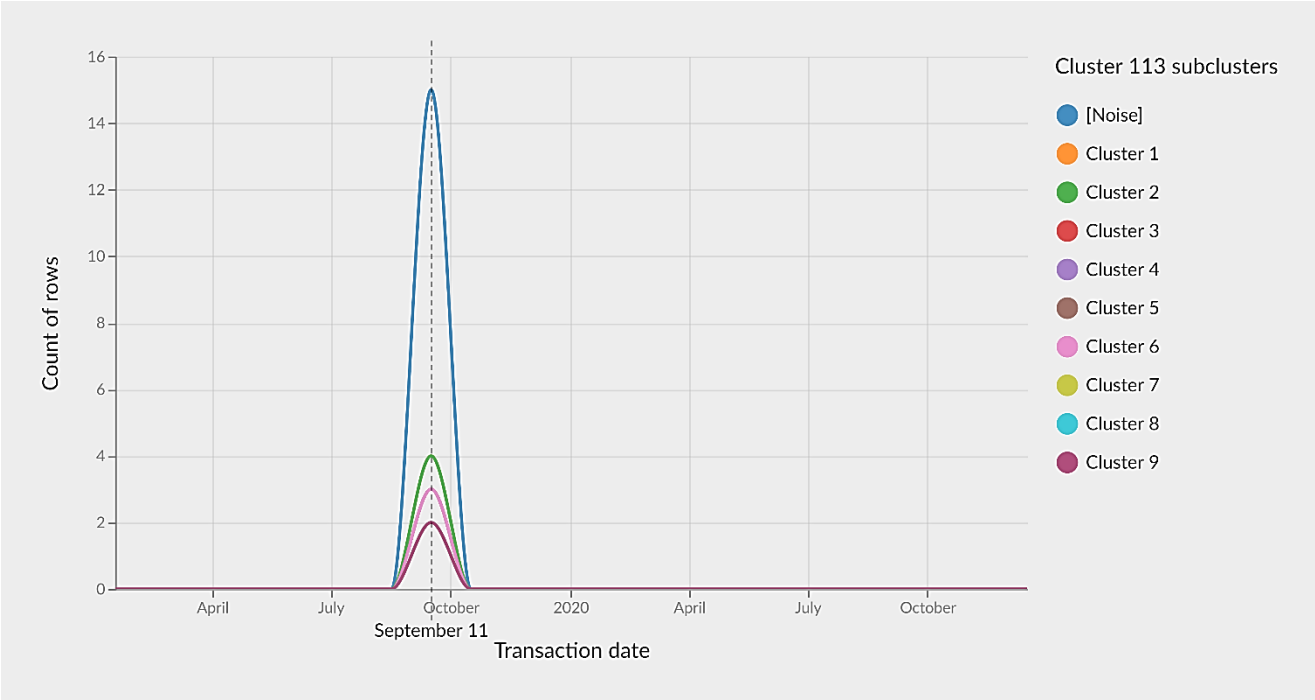


Figure 18. Transaction date of cluster 113 subclusters

Source: designed by author

The figure above shows all 9 subclusters and the noise of cluster 113 according to date – as we see, all the observations are from the same date – September 11, 2019.

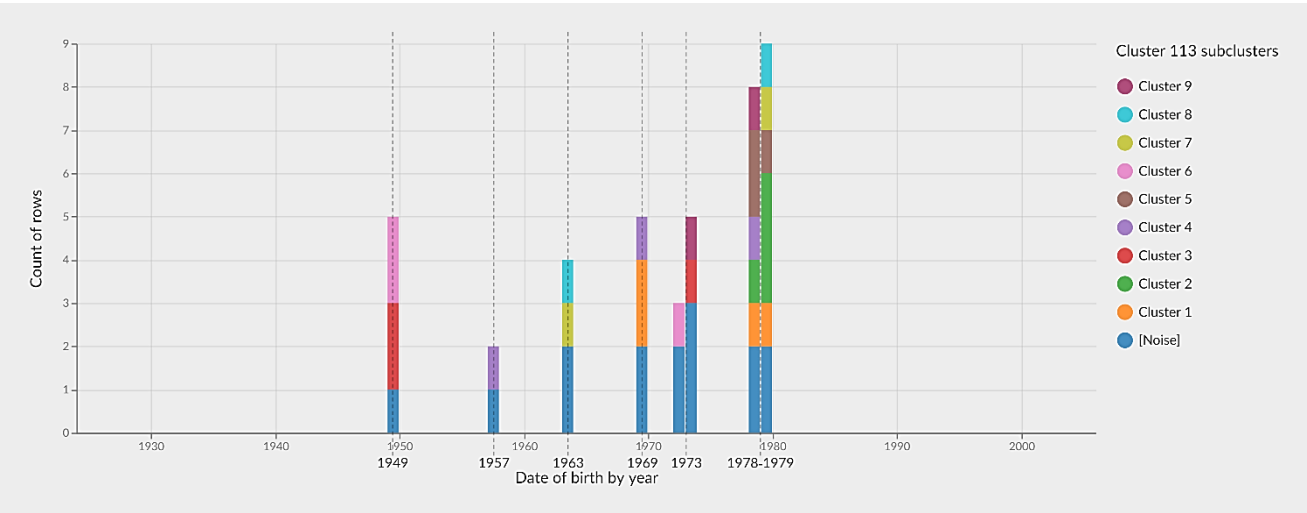


Figure 19. Customers' birth year across cluster 113 subclusters

Source: designed by author

Next, the the date of birth by year is examined in Figure 19. Overall, the people performing the fraudulent transactions in this group are born between 1949 and 1979, with the largest group consisting of those born at the end of this period, making them about 41 and 40 years old in 2019. Observations from subclusters 2 and 5 all fall within the the last years of this period only, while observations from other subclusters are move sperad out.

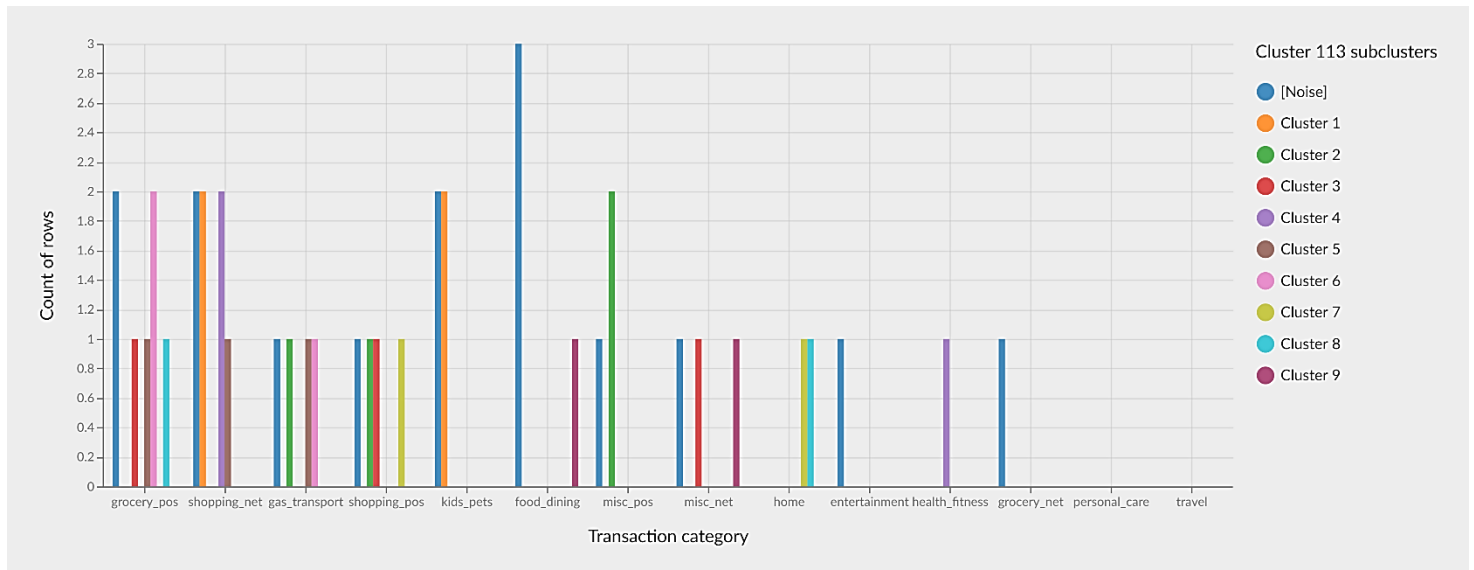


Figure 20. Transaction categories across cluster 113 subclusters

Source: designed by author

Figure 20 shows the subclusters of cluster 113 according to transaction category. The categories „grocery_pos“ and „shopping_net“ have observations from many different subclusters, in all encompassing 7 and 6 observations, respectively. The categories “gas_transport” and “shopping_pos” have less observations, but similar variety across subclusters as the first two categories. The category “kids_pets” has 2 observations in subclusters 1 and noise. The “food_dining” category captures 3 of the noise observations and 1 observation from subcluster 9. Other categories have less observations overall and less variety across different subclusters.

Exploring the merchants across different subclusters, no patterns were found, as observations were scattered across different merchants. The only merchant who was present more than once is “fraud_Killback LCC”, in the noise and 6th subclusters.

Additionally, I take a look at gender composition of the subclusters below, summarized in Figure 21.

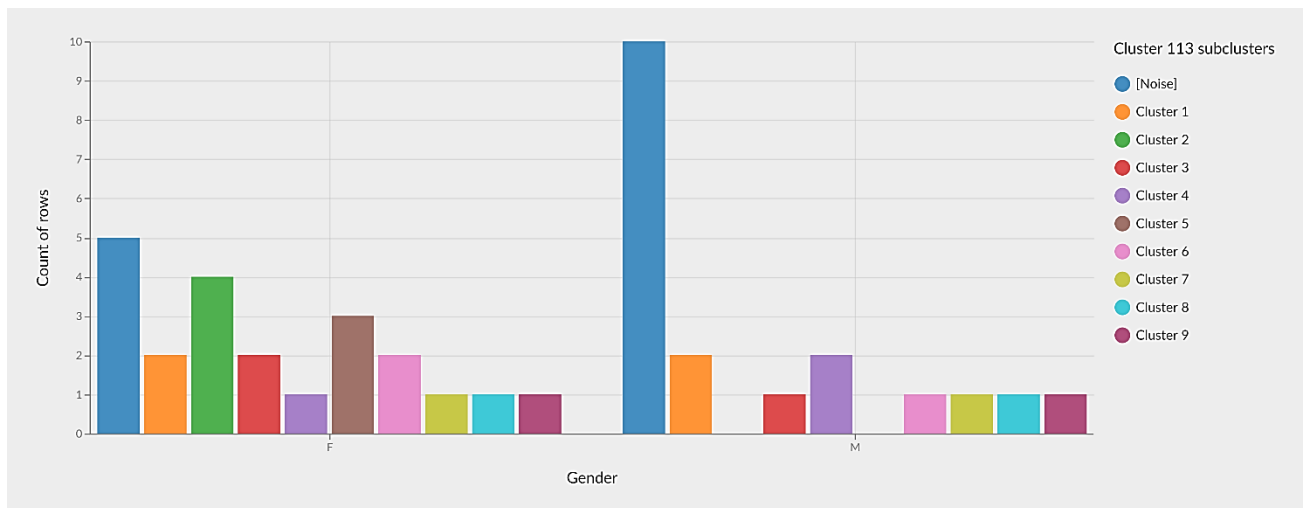


Figure 21. Observations according to gender across cluster 113 subclusters

Source: designed by author

From this figure we see that each of the subclusters has at least one observation of an illegitimate transaction attributable to women. Noise subcluster aside, the number of detected fraudulent transactions is greater for the female group in this case – 22 for females and 19 for males in all. Subclusters 2 and 5 consist of female observations only, while the last 3 subclusters have an even split between the two genders.

Next, I examine the subclusters of cluster 471, which had 35 fraudulent observations.

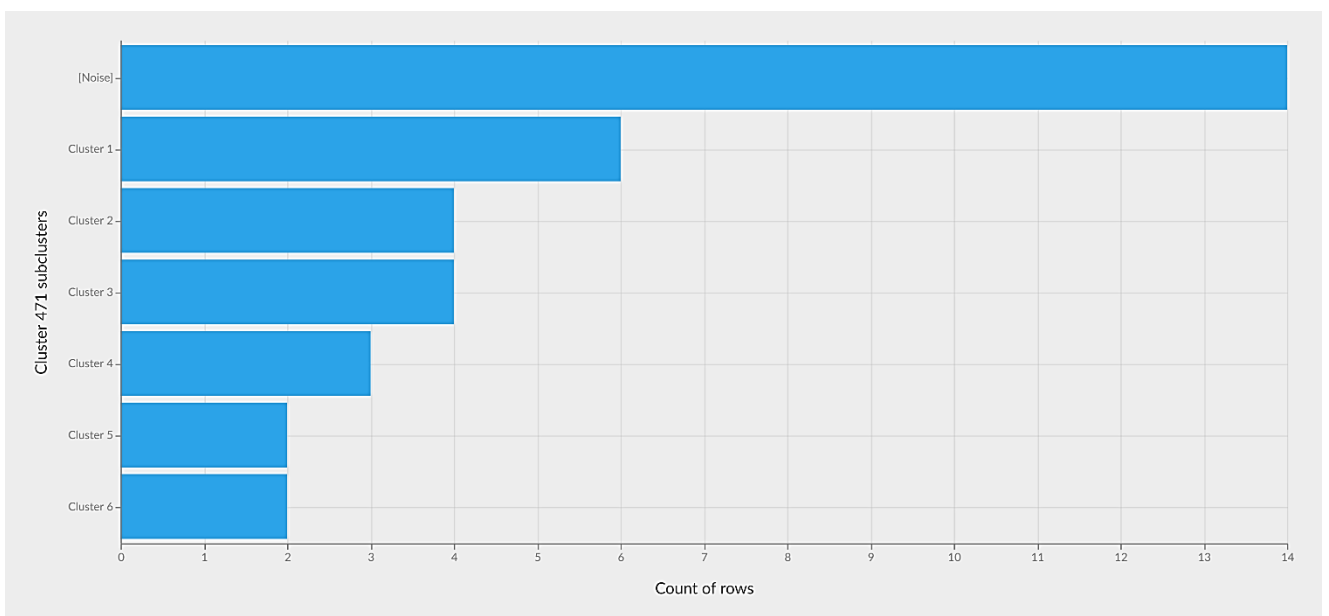


Figure 22. Subclusters of cluster 471

Source: designed by author

As we see from Figure 22, the fraudulent observations from the initial clusters have been divided into 6 subclusters and a subcluster catching the noise.

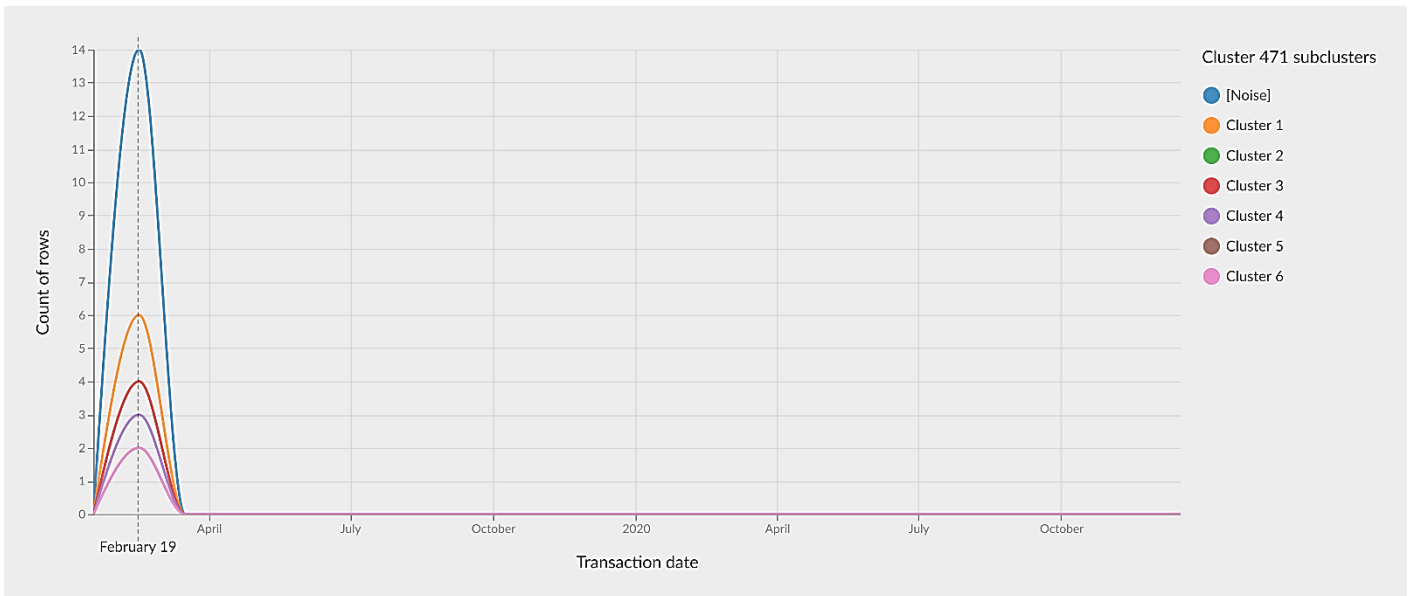


Figure 23. Transaction date of cluster 471 subclusters

Source: designed by author

Figure 23 presents the transaction date for the subclusters of cluster 471. We see that all of the fraudulent transactions have been performed on the same date – February 19th, 2019. This aligns with the findings in Figure 5, which showed that in 2019 from the year beginning to April the proportion of fraudulent transactions was growing.

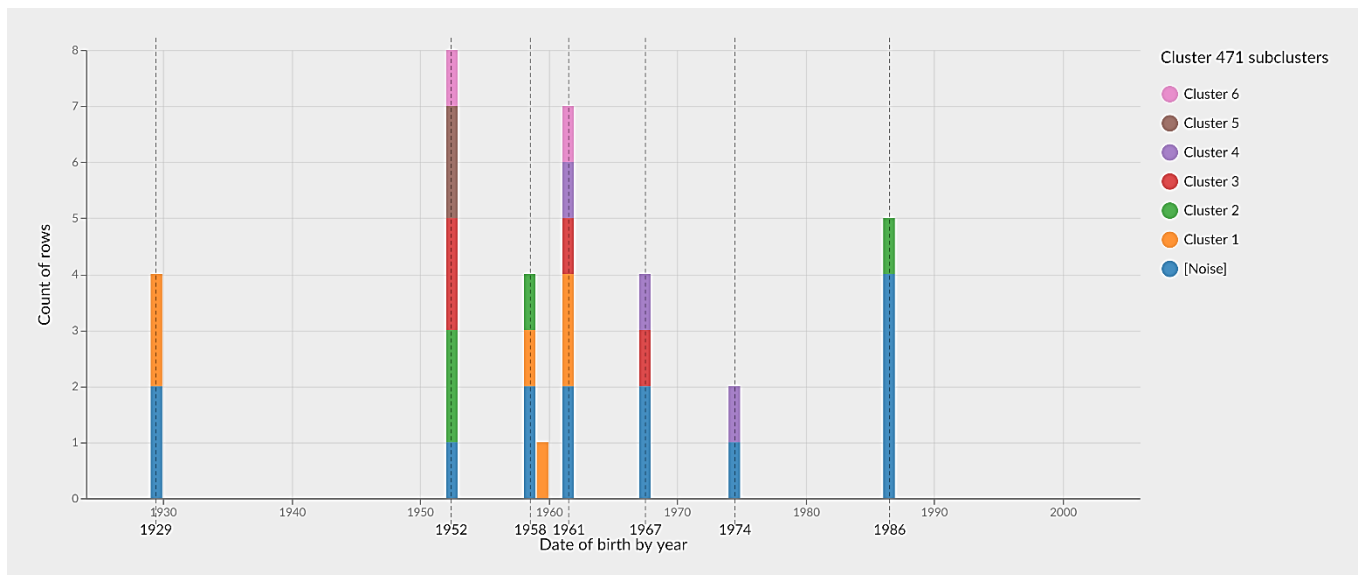


Figure 24. Customers' birth year across cluster 471 subclusters

Source: designed by author

Figure 24 shows in which year the people who performed the fraudulent transaction were born. Firstly, we observe that the year range in comparison to the subclusters generated from the cluster 113 is much wider – from 1929 to 1986. However, most observations are

concentrated in the period 1952 – 1961. 8 transactions are attributed to people who were born in 1952, which makes them around 67 years old in 2019, when the fraudulent transactions were performed. 7 fraudulent transactions were made by people born in 1961, who were 58 years old in 2019.

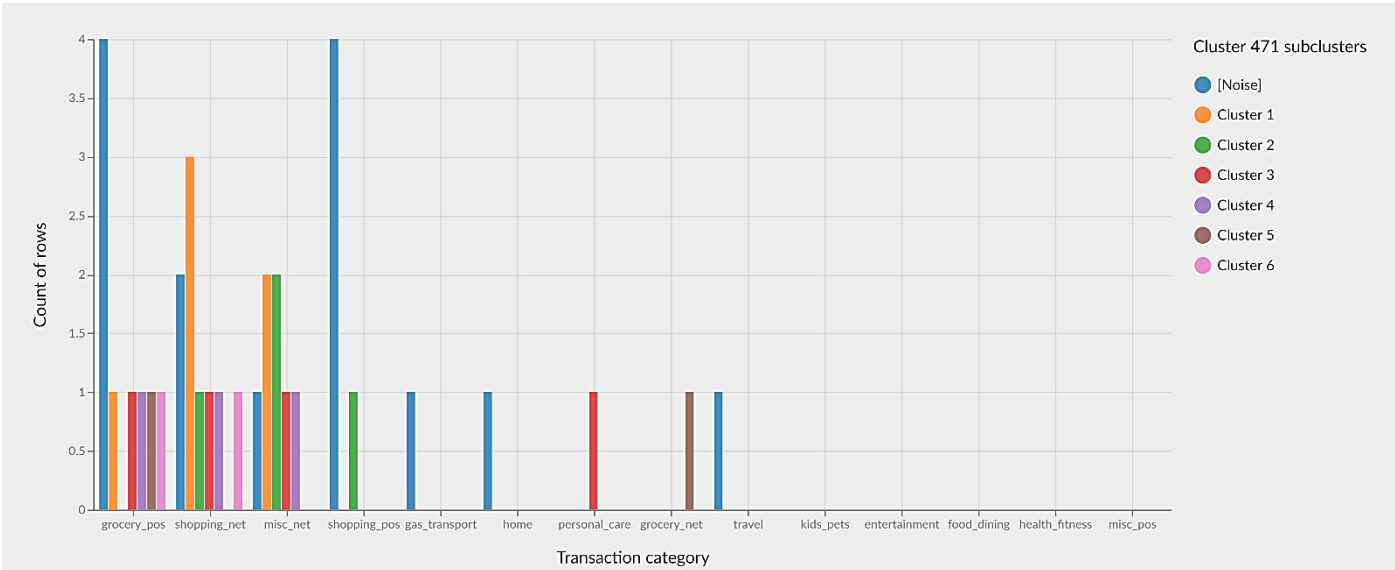


Figure 25. Transactions categories across cluster 471 subclusters

Source: designed by author

Next, the transaction category is analysed for the subclusters of 471 (Figure 25). Similarly to the case of 113 subclusters, most fraudulent observations are concentrated in the “grocery_pos” and “shopping_net” categories. Additionally, the category “misc_net” also captures quite a few observations from different subclusters, while categories like “shopping_pos”, “home” and “gas_transport” mostly capture the noise.

The analysis of the merchant did not yield any significant results as the observations are scattered between different merchants, like the case with cluster 113.

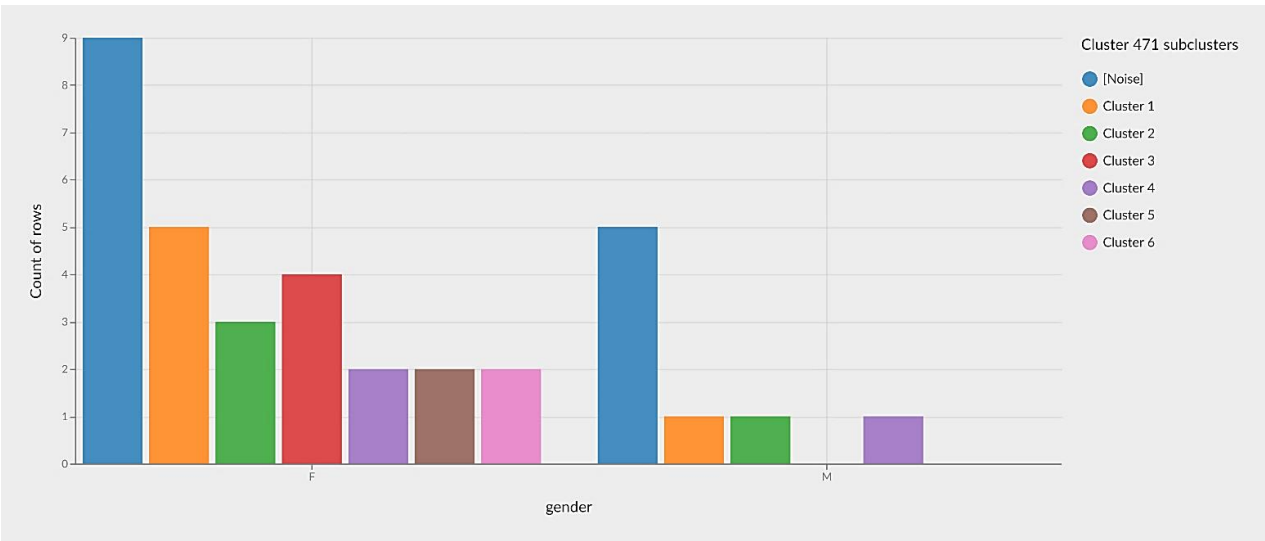


Figure 26. Observations according to gender across cluster 471 subclusters

Source: designed by author

Figure 26 presents the gender distribution between different subclusters of 471. Here, observations with women performing fraud outweigh the observations with men as fraudsters. This is similar to the situation with cluster 113 as well as the fact that there are some subclusters, like the 3rd, 5th and 6th where all of the transactions are performed by women.

Analysis of the generated subclusters of cluster 270 follows.

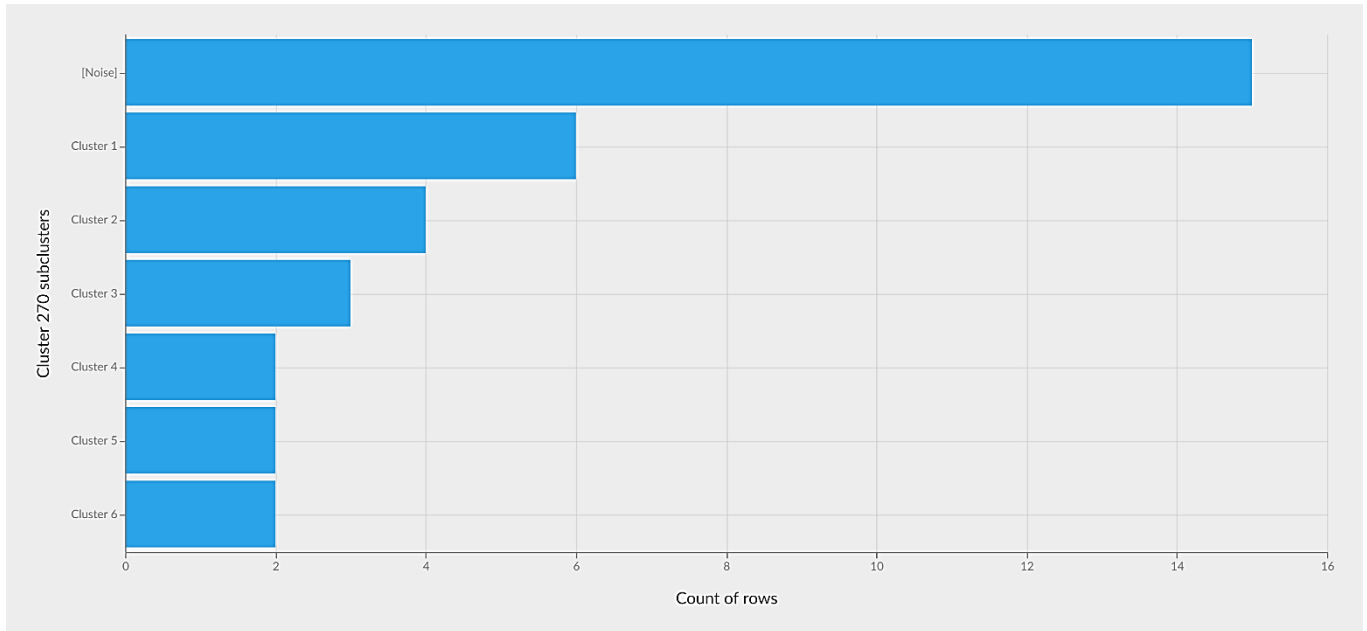


Figure 27. Subclusters of cluster 270

Source: designed by author

As per Figure 27, the 34 fraudulent transactions were allocated into 6 subclusters and a noise subcluster.

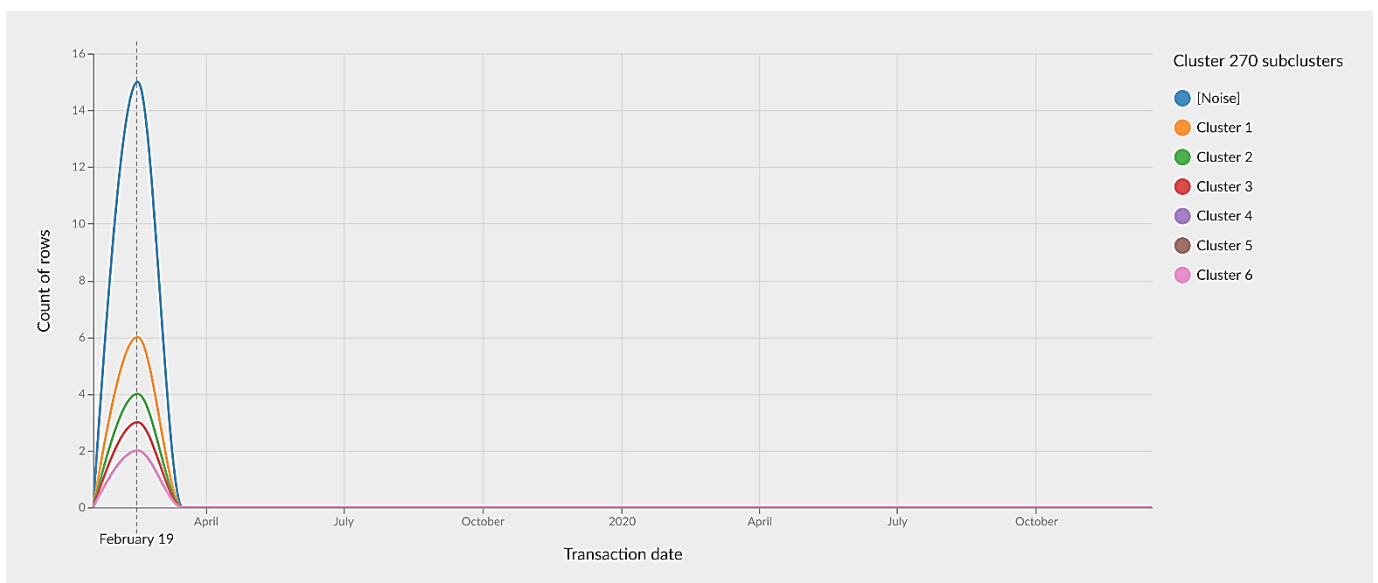


Figure 28. Transaction date of cluster 270 subclusters

Source: designed by author

Figure 28 shows the transaction date of fraudulent transactions of subclusters of 270th cluster – all of them were performed on February 19, 2019, the same date that the fraudulent transactions were performed in cluster 471.

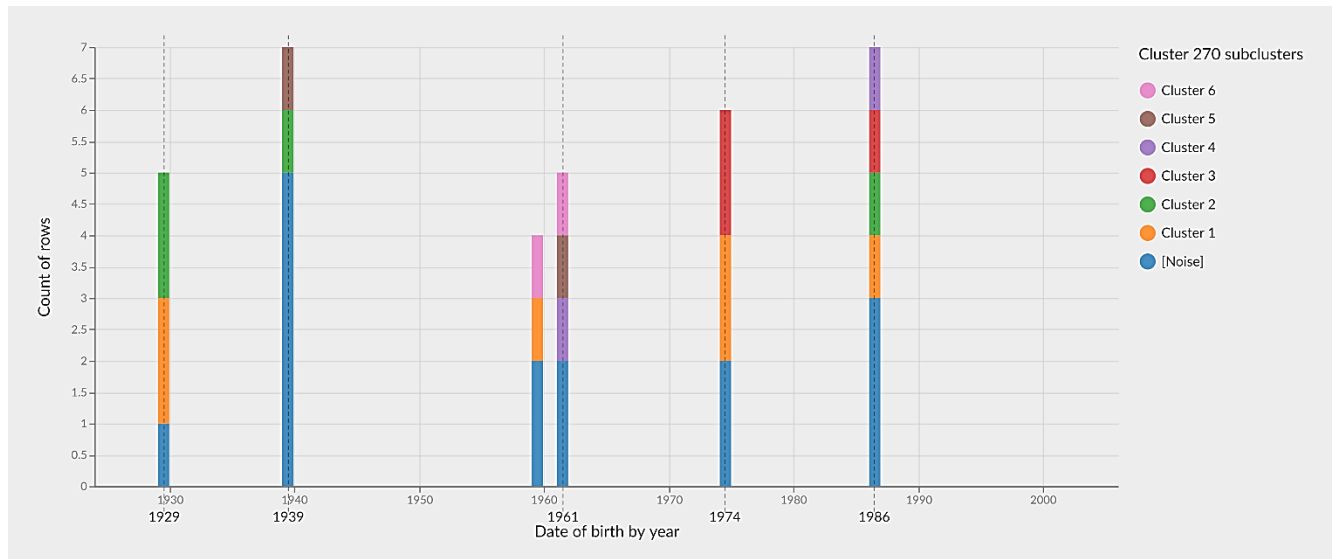


Figure 29. Customers' birth year across cluster 270 subclusters

Source: designed by author

Figure 29 presents year of birth of the people performing the fraudulent transactions across subclusters. In cluster 270, the range of the year of birth is the same as in cluster 471. However, in this case most of the fraudulent transactions were performed by people born in 1939 and 1986, making the people 80 and 33 years old in 2019.

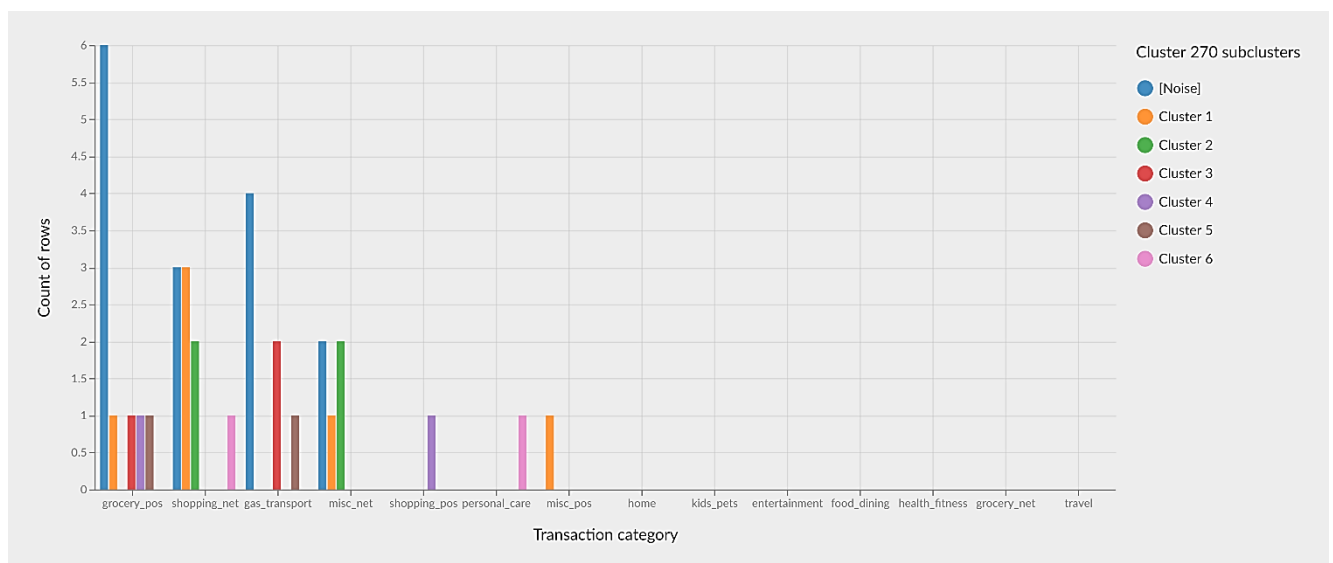


Figure 30. Transactions categories across cluster 270 subclusters

Source: designed by author

Figure 30 presents the transaction categories used in different transactions of subclusters of 270. Once again, we notice that categories “grocery_pos” and “shopping_net” have the most observations across different subclusters. Categories “gas_transport” and “misc_net” are secondary in term of relevance, having 7 and 5 observations across 3 subclusters.

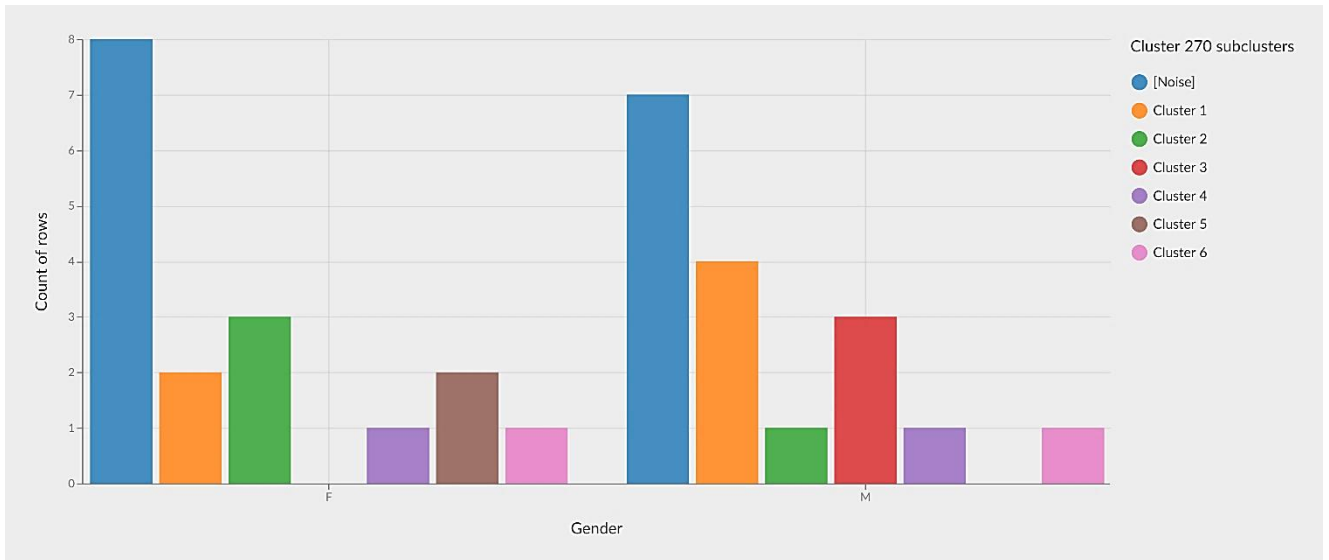


Figure 31. Observations according to gender across cluster 270 subclusters

Source: designed by author

Lastly, the subclusters are examined in terms of gender in Figure 31. Notice, that the distribution between genders in this group is much different than in case of the previous cluster and its subgroups – here we see that that in some clusters there are more observations relating to men, such as in subclusters 1 and 3, where the aforementioned subcluster consists solely of transactions made by men. However, subclusters 2 and 5 show the opposite scenario. In all, 17 observations are attributed to women and men, making the split half-in-half.

To sum up the cluster analysis, by generating subclusters from three clusters which have the most fraudulent transactions, some similar characteristics have been spotted. Firstly, all three clusters have most of their fraudulent transactions attributed to categories “grocery_pos” and “shopping_net”, other reoccurring categories include “gas_transport”, “shopping_pos” and “misc_net”. Interestingly, that the clusters were able to capture the first 5 categories shown in figure 8, where 4 of 5 had an overwhelming percentage of illegitimate transactions. Notice, that the fraudulent observations captured by clusters 471 and 270 were all made on February 19, when the proportion of fraudulent transactions was above 5 percent and was experiencing an increase. We also noticed that the age of the people performing the fraudulent transactions varies across clusters. From the three clusters examined and looking at the most common ages,

we can spot two groups of people in terms of age – a younger group, consisting of people aged 33, 40 and 41 and an older group, whose age include 58, 67 and 80 years. According to the findings of clusters 113 and 471 analysis, women more commonly perform fraudulent transactions, while the results from cluster 270 suggest a more even split between the genders.

3.4. Keyword analysis

To further examine the characteristics of fraudulent transactions, keyword analysis is performed – the text variable indicating the city is searched for specific keywords. For this part of the analysis only fraudulent transactions are considered, therefore allowing to pinpoint cities which are most associated with illegitimate transactions.

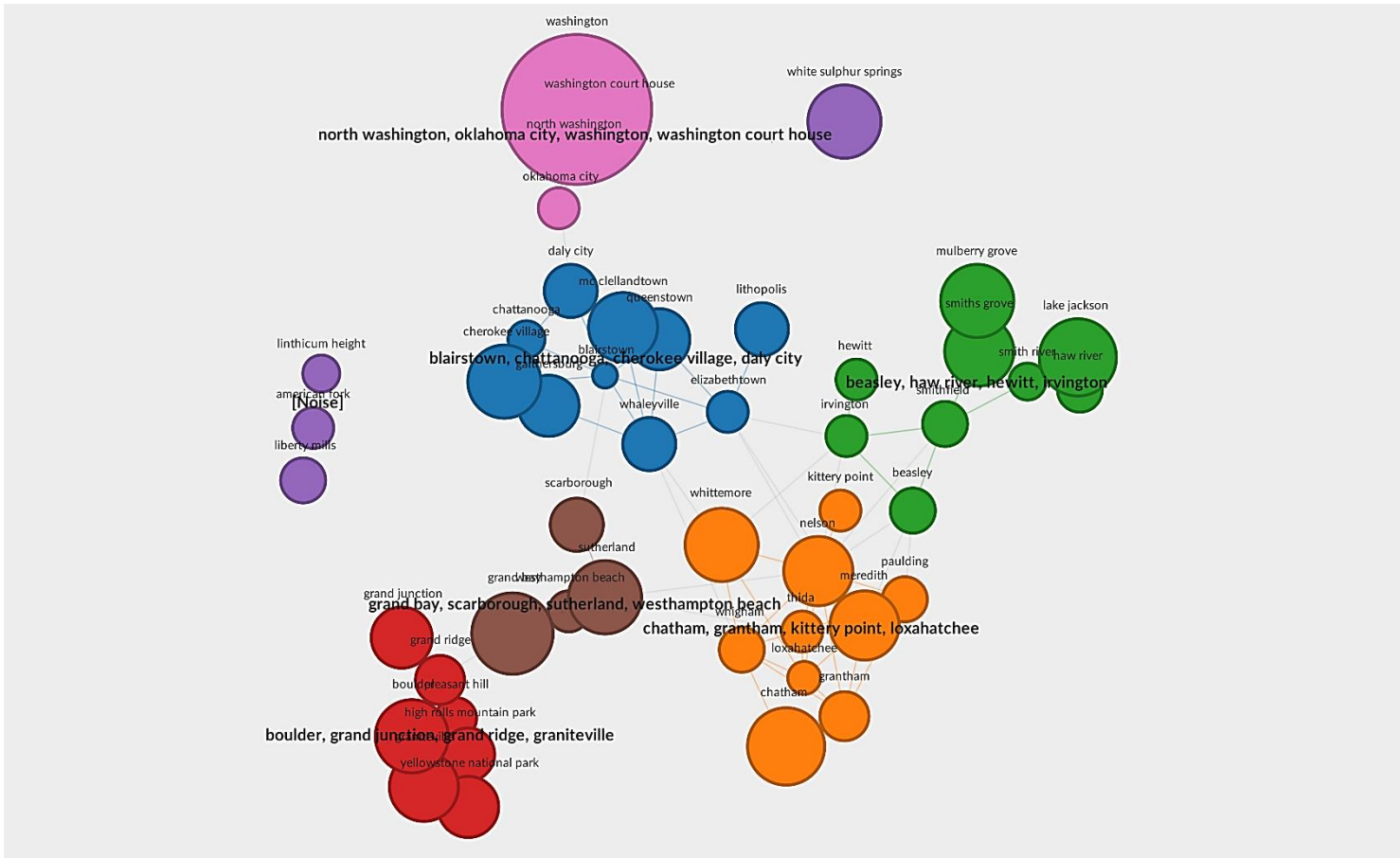


Figure 32. Keyword network visualization|

Source: designed by author

Graphext software has generated the following network visualization with the extracted keywords, pictured in Figure 32. We can spot seven clusters formed from the 48 keywords generated in total.

The figure below examines how many times certain words are repeated. This will be used to find the main keywords.

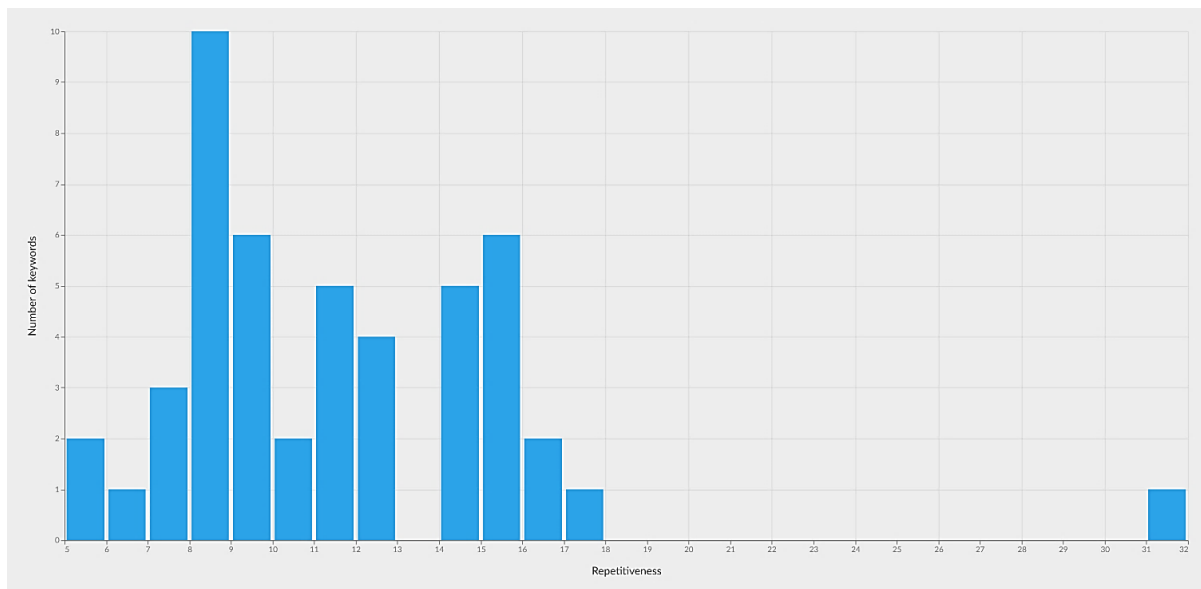


Figure 33. Reoccurrence of keywords within fraudulent transactions

Source: designed by author

Figure 33 suggests that at most, there are 10 keywords which were repeated 10 times. The repetition of the keyword ranges mostly from 5 to 18 times, however there is one keyword that was repeated 31 times throughout the fraudulent transactions – “Washington”. This can also be seen in the previous figure as the circle area representing this keyword is by far the largest. In terms of reoccurrence, there is also one keyword which was repeated 17 times and other two keywords repeated 16 times. As these keywords represent cities across United States, it is better to pin these locations on a map.

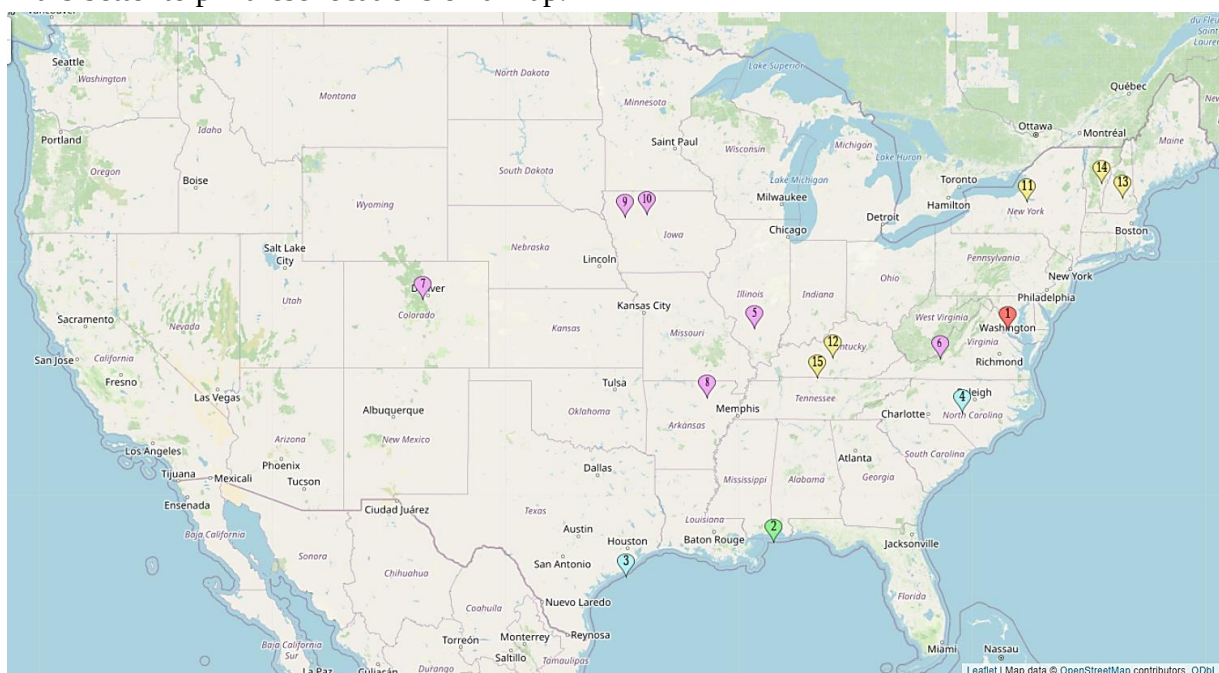


Figure 34. U.S.A map indicating cities with most association to fraudulent transactions

Source: designed by author using MapCustomizer (<https://www.mapcustomizer.com/>)

Figure 34 shows the locations representing the most common keywords on a map. The different colours of pins indicate the number how many times the location (or city) has been repeated – red (31), green (17), light blue (16), pink (15) and yellow (14). As we can see the majority of the cities listed are located on the east side of the country with the exception of Boulder. This suggests that cities located in this part of the country are more associated with fraudulent transactions, constituting a possible problem of increased fraud in the area. Interestingly, if we consider the first ten states which account for the biggest proportion of fraudulent transaction from Figure 9, most of them are located on the east side of the country as well, with the exception for California. Cities in the states of New York and Texas, which are first and second in terms of percentage of fraudulent transactions are present on this map.

Considering the definition of transaction fraud given by Sikkandar et al. (2023), the findings imply that there is a high concentration of fraudsters in these cities. As the people committing the crime with someone else's credit card either acquired and used it in the same city or area or moved to the location in order to perform the transaction perhaps with the aim of not being caught.

To conclude, this chapter has presented the execution and results of the study. Firstly, descriptive statistics were analyzed – the small proportion of fraudulent transactions constitutes a unbalanced dataset problem. Examining the transaction dates, we notice that the number of transactions was increased rapidly during year end of 2019 and 2020. Analysis of transaction amount showed that majority of them are between 0 to 3 000, maximum value of fraudulent transactions is 1376,04. Regarding gender, we notice that the percentage of fraudulent transactions committed by women is higher than that committed by men. Distribution of transaction category for fraudulent transactions states that 'grocery_pos' and 'shopping_net' are the most common, while the three US states with biggest proportion of illegitimate transactions were New York, Texas and Pennsylvania.

After the initial application of the classification techniques, training scores showed that Decision Tree had the highest accuracy rate. However, after parameter optimization, the accuracy of all the parameters has increased, however, in this stage K Nearest Neighbors had the highest score. The ROC curve results also showed that the K Nearest Neighbors classifier had the highest AUC score. Type I and II errors were calculated based on the confusion matrices, which implied that K Nearest Neighbors classification technique had the lowest type I error. Logistic Regression had the lowest type II error. When comparing the three main metrics – K nearest Neighbor classification technique was deemed as the best at detecting fraud, with nearly 90 percent precision and captured all of the illegitimate transactions. Decision Tree classifier had comparable results to the previous classification technique, although slightly

lower. Logistic regression showed the poorest results regarding fraud detection, as it was shown to be better at classification of legitimate transactions. Classification technique analysis could be of interest mainly for businesses and companies processing large quantities of transactions and looking into developing a fraud detection mechanism, which could possibly incorporate classification techniques as this tool is useful for separating the illegitimate transactions from others, while not providing many insights about the characteristics initially.

Cluster analysis presented the following characteristics of fraudulent transactions – most common transaction categories are “grocery_pos” and “shopping_net” as well as “gas_transport”, “shopping_pos” and “misc_net”. The fraudulent transactions are attributed to people, most commonly women, of different age groups – younger group of ages 33, 40 and 41 and the older group between the ages of 58 and 80. Keyword analysis was also helpful in distinguishing a key characteristic of fraudulent transactions in terms of location – by extracting keywords from the variable specifying the city where the transactions were performed, it was noticed that the city of Washington was the city associated the most with fraudulent transactions because of its repetition level. When analysing other keywords which were repeated 14 times or more, we observed that the cities where fraud was most prevalent were located on the east side of the country. Findings from the cluster and keyword analysis are important to study the patterns and overall characteristics of transaction fraud – these attributes can be used to study and pinpoint the location of fraudsters as well as the portrait of the victim, which can be used by the government and regulatory institutions to take action in lowering the fraud rate in the specified locations while keeping in mind the types of people most commonly falling victims to the fraud in order to offer protection or educational help.

CONCLUSIONS

1. From the literature review that was conducted the main types of financial fraud were described (credit card, mortgage, money laundering, financial statement, securities and commodities as well as insurance and cryptocurrency fraud) and methods of detecting financial fraud using Big Data (Support Vector Machine, Decision Tree, Random Forest, Logistic regression, Neural Network, Naïve Bayes, Social Network Analysis, ect.).

While reviewing the scientific articles gaps within the research realm was detected – certain types of financial fraud (mortgage, securaties and commodities as well as cryptocurrency fraud) have little to no scientific coverage. As well as different methods of detecting fraud were not utalized for all fraud categories mentioned.

A relevent problem in the field of financial fraud is the lack of available data due to privacy concerns.

2. To conduct this study three classification techniques were chosen - K Nearest Neighbors, Decision Tree and Logistic Regression. Decision Tree and Logistic Regression are commonly used in other scientific works and in addition to these two methods, a not so popular method - K Nearest Neighbors is utilized as of its reported great performance results in fraud detection. To compare the performance of mentioned classification techniques in terms of detecting fraud, three metrics commonly used by other researchers are calculated and compared - precision, recall, F-1 score and accuracy. In order to distinguish specific traits associated with fraudulent transactions two main tools were utilized – cluster and keyword analysis. Clustering groups transactions based on similarity with respect to a given variable, by focusing on the variable specifying fraud, we were able to examine the traits of those clusters comprised mostly of fraudulent cases and identify certain features. By using natural language processing, keywords are extracted from the data and analysed based on the frequency of repetition of a certain word, which is used to define further elements characterising fraudulent transactions.
3. After applying the classifiers on both training and test subsamples, it can be concluded that the K Nearest Neighbors classifier is the best performing method for fraud classification. It was able to capture all of the fraudulent transactions, as well as majority of the legitimate ones. Note, that the type II error was above 6 percent, which is important to consider when using this classificatory for fraud detection techniques. Decision Tree method is deemed as second best. Although its precision and recall metrics are lower, the results are still comparable to those of K Nearest Neighbors. Logistic Regression had the lowest scores overall, lacking behind in performance compared to the previous two methods. It was able

to correctly classify all of the non-fraudulent transactions and its type II error is 0, however, that is because of its inability to correctly classify a single fraudulent transaction as fraudulent. In terms of implementation, based on the current results, it would be best to implement the K Nearest Neighbors classifier in a fraud detection mechanism. However, one should keep in mind that these results are generated using synthetic data. For best and most accurate results, this study should be repeated using a particular company's transactional data to confirm the finding and adjust for the differences in transactional data. Keeping in mind the cost of the type II misclassification error, the company would be advised to conduct research of how costly in financial terms this error can be and further work in decreasing the error margin. Lastly, these results are applicable for credit card transactional data, as when results are compared for the same classifiers for detection of fraudulent cases in other types of fraud, results may differ, so considering different approaches and other classifiers might be best.

From the cluster and subcluster analysis the following characteristics of fraudulent transactions were defined – likely to be associated with categories relating to shopping, grocery as well as gas and transport, misc, age of the supposed people executing the transfer varies considerably, two age groups can be detected – younger groups of ages 33, 40 and 41 and older group between the ages of 58 and 80 years. In terms of gender, women were found to have performed more fraudulent transactions, although this finding varies across clusters.

Keyword analysis provided evidence that the cities most associated with fraudulent transactions are located on the east side of United States, where Washington was the most repeated keyword across fraudulent transactions. This implies that the people committing the crime are concentrated or are often in these specific cities or areas. This is useful information for regulatory and local government authorities who can work together with companies to develop fraud prevention mechanisms and take insights relating to the types of people mostly affected by transaction fraud to take regulatory and educational measures to reduce the fraud rate.

As mentioned in the introduction the main beneficiaries of the findings of this research are businesses, financial institutions as well as government and regulatory authorities. The information regarding characteristics of the fraudulent transactions such as gender, age, transaction category and city can be used by local governments to create a profile of people most affected by transaction fraud and take regulatory and educational measures to prevent fraud cases in the specific areas and cities.

REFERENCES

- Aljassmi, M., Gamal, A. a. M., Jalil, N. A., & Viswanathan, K. K. (2023). An analysis of the determinants of money laundering in the United Arab Emirates (UAE). *Journal of Money Laundering Control*. <https://doi.org/10.1108/jmlc-09-2023-0150>
- Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019 [Article]. *Computer Science Review*, 40, 100402. <https://doi.org/10.1016/j.cosrev.2021.100402>
- Amponsah, A. A., Adekoya, A. F., & Weyori, B. A. (2022). A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology. *Decision Analytics Journal*, 4, 100122. <https://doi.org/10.1016/J.DAJOUR.2022.100122>
- Antwi, S., Tetteh, A. B., Armah, P., & Dankwah, E. O. (2023). Anti-money laundering measures and financial sector development: Empirical evidence from Africa. *Cogent Economics & Finance*, 11(1). <https://doi.org/10.1080/23322039.2023.2209957>
- Association of Certified Fraud Examiners (ACFE). (2022). *Occupational Fraud 2022: A Report to the Nations*.
- Attigeri G., Manohara M. M. Pai, & Radhika M Pai. (2021). Supervised Models for Loan Fraud Analysis using Big Data Approach [Article]. *Engineering Letters*, 29(4), 1422.
- Baumer, E. P., Arnio, A. N., & Wolff, K. T. (2013). Assessing the role of mortgage fraud, confluence, and spillover in the contemporary foreclosure crisis. *Housing Policy Debate*, 23(2), 299–327. <https://doi.org/10.1080/10511482.2012.727843>
- Baumer, E. P., Ranson, J. W. A., Arnio, A. N., Fulmer, A., & De Zilwa, S. (2017). Illuminating a Dark Side of the American Dream: Assessing the Prevalence and Predictors of Mortgage Fraud across U.S. Counties. *American Journal of Sociology*, 123(2), 549–603. <https://doi.org/10.1086/692719>
- Bhowmik, A., Sannigrahi, M., Chowdhury, D., Dwivedi, A. D., & Rao Mukkamala, R. (2022, December 17). DBNex: Deep Belief Network and Explainable AI based Financial Fraud Detection. 2022 IEEE International Conference on Big Data (Big Data). <https://doi.org/10.1109/bigdata55660.2022.10020494>
- Bi, W., & Liang, Y. (2022). *Risk Assessment of Operator's Big Data Internet of Things Credit Financial Management Based on Machine Learning*. <https://doi.org/10.1155/2022/5346995>
- Bologa, A.-R., Bologa, R., & Florea, A. (2013). Big Data and Specific Analysis Methods for Insurance Fraud Detection. [Article]. *Database Systems Journal*, 4(4), 30–40.

- Boulieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2023). Fraud detection with natural language processing. *Machine Learning*. <https://doi.org/10.1007/s10994-023-06354-5>
- Chen, J., Tao, Y., Wang, H., & Chen, T. (2015). Big data based fraud risk management at Alibaba [Article]. *The Journal of Finance and Data Science*, 1(1), 1–10. <https://doi.org/10.1016/j.jfds.2015.03.001>
- Chen, Z., Van Khoa, L., Teoh, E. N., Nazir, A., Karupiah, E. K., & Lam, K. H. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, 57(2), 245–285. <https://doi.org/10.1007/s10115-017-1144-z>
- Cheng, C.-H., Kao, Y.-F., & Lin, H.-P. (2021). A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes [Article]. *Applied Soft Computing*, 108, 107487. <https://doi.org/10.1016/j.asoc.2021.107487>
- Cong, L. W., Li, X., Tang, K., & Yang, Y. (2022). Crypto Wash trading. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4312030>
- Cumming, D., Leung, T. Y., & Rui, O. (2015). GENDER DIVERSITY AND SECURITIES FRAUD. *Academy of Management Journal*, 58(5), 1572–1593. <https://doi.org/10.5465/amj.2013.0750>
- Domashova, J., & Mikhailina, N. (2021). Usage of machine learning methods for early detection of money laundering schemes. *Procedia Computer Science*, 190, 184–192. <https://doi.org/10.1016/J.PROCS.2021.06.033>
- Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging Financial Social Media Data for Corporate Fraud Detection [Article]. *Journal of Management Information Systems*, 35(2), 461–487. <https://doi.org/10.1080/07421222.2018.1451954>
- Dong, W., Moses, C., & Li, K. (2011). Efficient K-nearest neighbor graph construction for generic similarity measures. *Proceedings of the 20th International Conference on World Wide Web*. <https://doi.org/10.1145/1963405.1963487>
- Dong, Y., Jiang, Z., Alazab, M., & Kumar, P. M. (2021). Real-time fraud detection in e-market using machine learning algorithms. *Journal of Multiple-Valued Logic and Soft Computing*, 36(1-3), 191-209.
- Durguti, E., Arifi, E., Gashi, E., & Spahiu, M. (2023). Anti-money laundering regulations' effectiveness in ensuring banking sector stability: Evidence of Western Balkan. *Cogent Economics & Finance*, 11(1). <https://doi.org/10.1080/23322039.2023.2167356>
- Elul, R., Payne, A., & Tilson, S. (2023). Owner-occupancy fraud and mortgage performance. *Real Estate Economics*, 51(5), 1137–1177. <https://doi.org/10.1111/1540-6229.12455>

- Federal Reserve. (2023). *Federal Reserve Payments Study (FRPS)*.
<https://www.federalreserve.gov/paymentsystems/fr-payments-study.htm>
- Fiederling, K., Schiller, J., & von Bieberstein, F. (2018). Can we Trust Consumers' Survey Answers when Dealing with Insurance Fraud? *Schmalenbach Business Review*, 70(2), 111–147. <https://doi.org/10.1007/s41464-017-0041-z>
- Fronzetti Colladon, A., & Remondi, E. (2017). Using social network analysis to prevent money laundering [Article]. *Expert Systems with Applications*, 67, 49–58.
<https://doi.org/10.1016/j.eswa.2016.09.029>
- Gepp, A., Kumar, K., & Bhattacharya, S. (2023). Taking the hunch out of the crunch: A framework to improve variable selection in models to detect financial statement fraud. *Accounting & Finance*. <https://doi.org/10.1111/acfi.13192>
- Graphext, An Introduction to Natural Language Processing for Text Analysis. (n.d.).
<https://www.graphext.com/docs/an-introduction-to-natural-language-processing-for-text-analysis>
- Graphext, Keywords. (n.d.). <https://www.graphext.com/docs/analysis-types-text>
- Graphext, Models. (n.d.). <https://www.graphext.com/docs/analysis-types-models>
- Hafiz, K. T., Aghili S. and Zavorsky P. (2016). The use of predictive analytics technology to detect credit card fraud in Canada. 2016 11th *Iberian Conference on Information Systems and Technologies (CISTI)*, Gran Canaria, Spain, 2016, pp. 1-6,
<https://doi.org/10.1109/CISTI.2016.7521522>
- How UMAP Works — umap 0.5 documentation. (n.d.). https://umap-learn.readthedocs.io/en/latest/how_umap_works.html
- Islam, M. A., Uddin, M. A., Aryal, S., & Stea, G. (2023). An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes. *Journal of Information Security and Applications*, 78, 103618.
<https://doi.org/10.1016/j.jisa.2023.103618>
- Jain, H. (2022). Advanced Big Data Analysis Approach for Credit Card Fraud Detection. [Article]. *Grenze International Journal of Engineering & Technology (GIJET)*, 8(2), 1048–1054.
- Jan, C.-L. (2021). Detection of Financial Statement Fraud Using Deep Learning for Sustainable Development of Capital Markets under Information Asymmetry [Article]. *Sustainability (Basel, Switzerland)*, 13(17), 9879. <https://doi.org/10.3390/su13179879>
- Kaituko, L. E., Githaiga, P. N., & Chelogoi, S. K. (2023). Board structure and the likelihood of financial statement fraud. Does audit fee matter? Evidence from manufacturing firms

- in the East Africa community. *Cogent Business & Management*, 10(2).
<https://doi.org/10.1080/23311975.2023.2218175>
- Kerr, D., Loveland, K. A., Smith, K. T., & Smith, L. M. (2023). Cryptocurrency risks, fraud cases, and financial performance. *Risks*, 11(3), 51. <https://doi.org/10.3390/risks11030051>
- Leite, R. A., Gschwandtner, T., Miksch, S., Gstrein, E., & Kuntner, J. (2020). NEVA: Visual Analytics to identify Fraudulent Networks. *Computer Graphics Forum*, 39(6), 344–359.
<https://doi.org/10.1111/cgf.14042>
- Li, J. (2022). *E-Commerce Fraud Detection Model by Computer Artificial Intelligence Data Mining*. <https://doi.org/10.1155/2022/8783783>
- Madhurya, M. J., Gururaj, H. L., Soundarya, B. C., Vidyashree, K. P., & Rajendra, A. B. (2022). Exploratory analysis of credit card fraud detection using machine learning techniques. *Global Transitions Proceedings*, 3(1), 31–37.
<https://doi.org/10.1016/J.GLTP.2022.04.006>
- Marais, A., Vermaak, C., & Shewell, P. (2023). Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models. *Cogent Economics & Finance*, 11(1). <https://doi.org/10.1080/23322039.2023.2190215>
- Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011, February). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569.
<https://doi.org/10.1016/j.dss.2010.08.006>
- Nghiem, H., Murić, G., Morstatter, F., & Ferrara, E. (2021). Detecting cryptocurrency pump-and-dump frauds using market and social signals. *Expert Systems With Applications*, 182, 115284. <https://doi.org/10.1016/j.eswa.2021.115284>
- Nguyen, T. H., & Pontell, H. N. (2010). Mortgage origination fraud and the global economic crisis. *Criminology & Public Policy*, 9(3), 591–612.
<https://doi.org/10.1111/j.1745-9133.2010.00653.x>
- Patsakis, C., Πολίτου, E., Alepis, E., & Hernández-Castro, J. (2023). Cashing out crypto: state of practice in ransom payments. *International Journal of Information Security*.
<https://doi.org/10.1007/s10207-023-00766-z>
- Payment System Regulator. (2023). *PSR confirms new requirements for APP fraud reimbursement*.
- PwC. (2022). *PwC's Global Economic Crime and Fraud Survey 2022. Protecting the perimeter: The rise of external fraud*.
- Ranson, J. W. A., Arnio, A. N., & Baumer, E. P. (2018). Extending Research on Neighborhoods and Crime: An Examination of Mortgage Fraud across Chicago Census

- Tracts. *Journal of Quantitative Criminology*, 35(3), 465–491.
<https://doi.org/10.1007/s10940-018-9392-y>
- Ryan, E. (2023). Save the kids: The need for regulation of cryptocurrency to protect adolescents from fraud. *Family Court Review*, 61(3), 650–664.
<https://doi.org/10.1111/fcre.12729>
- Saia, R., & Carta, S. (2019). Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks [Article]. *Future Generation Computer Systems*, 93, 18–32. <https://doi.org/10.1016/j.future.2018.10.016>
- Setiawan, R., Tjahjono, B., Firmansyah, G., & Akbar, H. (2023). Fraud Detection in Credit Card Transactions Using HDBSCAN, UMAP and SMOTE Methods. *International Journal of Science, Technology & Management*, 4(5), 1333-1339.
<https://doi.org/10.46729/ijstm.v4i5.929>
- Shabbir, A., Shabir, M., Javed, A. R., Chakraborty, C., & Rizwan, M. (2022). Suspicious transaction detection in banking cyber–physical systems [Article]. *Computers & Electrical Engineering*, 97, 107596. <https://doi.org/10.1016/j.compeleceng.2021.107596>
- Shen, Y., Guo, C., Li, H., Chen, J., Guo, Y., & Qiu, X. (2021). Financial Feature Embedding with Knowledge Representation Learning for Financial Statement Fraud Detection. *Procedia Computer Science*, 187, 420–425.
<https://doi.org/10.1016/J.PROCS.2021.04.110>
- Sikkandar, Haseena & Subbaraj, Saroja & N, Suseandhiran & B, Manikandan. (2023). An intelligent approach for anomaly detection in credit card data using bat optimization algorithm. *Inteligencia Artificial*. 26. 202-222.
<https://doi.org/10.4114/intartif.vol26iss72pp202-222>
- Singh, K., & Best, P. (2019). Anti-Money Laundering: Using data visualization to identify suspicious activity [Article]. *International Journal of Accounting Information Systems*, 34, 100418. <https://doi.org/10.1016/j.accinf.2019.06.001>
- Sirisha Madhuri, T., Ramesh Babu, E., Uma, B., & Muni Lakshmi, B. (2021). Big-data driven approaches in materials science for real-time detection and prevention of fraud. *Materials Today: Proceedings*. <https://doi.org/10.1016/J.MATPR.2021.04.323>
- Steffensmeier, D. J., Harris, C. T., & Painter-Davis, N. (2015). Gender and arrests for larceny, fraud, forgery, and embezzlement: conventional or occupational property crime offenders? *Journal of Criminal Justice*, 43(3), 205–217.
<https://doi.org/10.1016/j.jcrimjus.2015.03.004>

- Warren, D. E., & Schweitzer, M. E. (2018). When Lying Does Not Pay: How Experts Detect Insurance Fraud. *Journal of Business Ethics*, 150(3), 711–726.
<https://doi.org/10.1007/s10551-016-3124-8>
- West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66. <https://doi.org/10.1016/J.COSE.2015.09.005>
- Yan, C., Li, Y., Liu, W., Li, M., Chen, J., & Wang, L. (2020). An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification [Article]. *Neurocomputing (Amsterdam)*, 393, 115–125.
<https://doi.org/10.1016/j.neucom.2017.12.072>
- Yu, X. (2013). Securities Fraud and Corporate Finance: Recent Developments. *Managerial and Decision Economics*, 34(7-8), 439–450. <https://doi.org/10.1002/mde.2621>
- Zhang, J., Fei, J., Song, X., & Feng, J. (2021). An Improved Louvain Algorithm for Community Detection. *Mathematical Problems in Engineering*.
<https://doi.org/10.1155/2021/1485592>
- Zhang, Y., Hu, A., Wang, J., & Zhang, Y. (2022). Detection of fraud statement based on word vector: Evidence from financial companies in China [Article]. *Finance Research Letters*, 46, 102477. <https://doi.org/10.1016/j.frl.2021.102477>

APPENDICES

Appendix A. Scientific paper published in the research area of the thesis.

FACTORS INFLUENCING FRAUDULENT TRANSACTIONS FROM BIG DATA PERSPECTIVE

Fabiana LEVON², Nijolė MAKNICKIENĖ 

^{1,2}*Department of Financial Engineering, Faculty of Business Management, Vilnius Tech University, Saulėtekio av. 11, 10223, Vilnius, Lithuania*
Department and/or Faculty, University, Address, City, Country

Received day Month 2023; accepted day Month 2023

Abstract. This article focuses on fraudulent behaviour and patterns as well as ways of detecting such patterns by using Big Data. The study analyses scientific articles to examine types of financial fraud and their detection techniques as well as develops a model that is based on factors characterizing fraudulent credit card transactions made across USA. Regression analysis, correlation and descriptive statistics analysis is applied. Statistically significant results are found indicating a causal relationship between fraudulent transactions and transactions made in Alaska, during the month of October and on a Thursday. Although, the impact of these relationships is relatively small. Expanding the dataset with more numerical variables that could be used for identifying fraudulent transactions is advised for future research as to better the overall fit of the model.

Keywords: Financial fraud, fraud detection, big data analytics, credit card transaction fraud, fraud detection methods.

JEL Classification: G20/G29/G28

Introduction

Financial fraud is a widely spread problem that affects both consumer and company or business – according to the results of PwC's (2022) Global Economic Crime and Fraud Survey of 2022, "46% of surveyed organizations reported experiencing fraud, corruption or other economic crimes in the last 24 months". Furthermore, the losses that come as a direct result of financial fraud can be very significant, as organizations lose 5 percent of revenue each year are attributed to fraud, which adds up to more than 3.6 billion of US dollars – as reported by the Association of Certified Fraud Examiners (ACFE) (2022). In terms of indirect costs, the companies face the risk of losing current and future customers – as major fraud incidents build upon the distrust between customer and company, affecting the business' reputation. Therefore, companies are determined to combat this issue with the help of fraud detection mechanisms. Thanks to today's advancements in technology, Big Data analytics can be utilized for this task in many forms. Not only does it use the copious amounts of data available in the online and electronic platforms to

it's advantage, but it is also considerably more reliable and cheaper than manual labour.

Financial fraud in itself is very broad and can be grouped into certain types, which are presented and discussed in this paper. For the purposes of clarity, this article focuses specifically on analysing fraudulent credit card transactions and what characteristic factors and patterns can be used to detect them. The novelty aspect of this study is that the findings can be used to build a system that could identify possible fraudulent transactions.

Problem – Prevalence of financial fraud in financial institutions.

Research object – Fraudulent behaviour and patterns.

Objectives – to find characteristics of transactional data which can help to determine a fraudulent transaction.

Purpose – To establish a model that could find fraudulent transactions by their characteristics.

This paper is composed of the following parts – firstly, the existing literature of the topic is

² * Corresponding author. E-mail: fabiana.levon@stud.vilniustech.lt

examined, then the methodology used in this paper is presented. The section discussing the results follows and the paper is finalized by providing conclusions.

1. Theoretical background

1.1. Types of financial fraud

Financial fraud - a broad term referring to financial gain accumulated via the intentional use of illegal means (West & Bhattacharya, 2016). To comprehend what exactly financial fraud is and how it is executed several types of financial fraud can be distinguished. Firstly, West & Bhattacharya (2016) defined three general types of financial fraud – bank fraud, corporate fraud and insurance fraud while composing a review of scientific articles on financial fraud detection spanning from 2004 to 2014. Similarly, by Al-Hashedi & Magalingam (2021) also make use of the same general financial fraud types in their overview of data mining techniques applied for financial fraud detection, while adding cryptocurrency fraud as a new type of fraud.

Each general fraud group is later split into more defined categories: bank fraud is specified as credit card fraud, mortgage fraud and money laundering fraud, corporate fraud encompasses financial statements fraud and securities and commodities fraud, lastly, healthcare and automobile insurance fraud both fall under the insurance fraud type. (West & Bhattacharya, 2016). In addition to the previously mentioned categories, Al-Hashedi & Magalingam (2021) also mention Bitcoin fraud as part of cryptocurrency related fraud.

Both, West & Bhattacharya (2016)) and Al-Hashedi & Magalingam (2021) define similar categories of financial fraud. Both works consider insurance fraud as a one category and do not split it when examining the different types. In terms of differences, Al-Hashedi & Magalingam (2021) adds to existing categories by introducing cryptocurrency fraud and including the associate the securities and commodities fraud category in both insurance and corporate fraud types.

1. Credit card fraud

Credit card fraud is described as illegal use of a one's credit card for the purpose of conducting fraudulent transactions without the consent of the credit card's owner (West & Bhattacharya, 2016). Al-Hashedi & Magalingam (2021) distinguish two kinds of credit card fraud – Online and Offline. In the latter the fraudsters use the physical credit card to perform illegal transactions without the owner's consent. Online

fraud entails remotely committed fraud, by using the Internet or electronic devices. As Fin Tech sector expands with more payment options available, the fraudsters are quick to exploit them. Credit card service providers who do not have a reliable fraud monitoring system are at risk of suffering financial and reputational losses (Hafiz et al., 2016).

2. Mortgage fraud

Mortgage fraud entails criminals targeting mortgage documents and manipulating or removing information in these documents during the process of loan application (Al-Hashedi & Magalingam, 2021). As a result the value of the property may be affected and misinterpreted, influencing the lender's decision on funding the loan (West & Bhattacharya, 2016).

3. Money laundering fraud

Money laundering is used to place illegally gained money into valid businesses. Concealing the origin of the money, it gives off the false appearance of legal income (West & Bhattacharya, 2016). This makes tracking down the funds and criminals very difficult. To add, the funds are often used to commit other crimes, such as funding terrorists and weapon trading (Al-Hashedi & Magalingam, 2021). Often the case, when there is lack of information about a certain organization's transactions to concur whether the organization is prone to money laundering (Domashova & Mikhailina, 2021).

4. Financial statement fraud

Financial statements include sensitive information regarding a business such as earnings, loans and other. Additionally, employee statements about the company's image and financial situation can be included (Al-Hashedi & Magalingam, 2021). Financial statement fraud seeks to modify these documents to make the company seem more profitable, have a more favourable status overall (West & Bhattacharya, 2016). This can affect the company's stock price, tax obligations or give off the impression of a well-functioning business to appease management. As this type of fraud is typically carried out by employees who have extensive knowledge about the documents it is difficult to identify (West & Bhattacharya, 2016).

5. Securities and commodities fraud

Popular examples of this type of fraud include embezzlement, Pyramid and Ponzi Schemes. These tactics use false information to trick the person into investing money in a business or company. (West & Bhattacharya, 2016).

6. Insurance fraud

Insurance fraud takes advantage of insurance policies meant to protect businesses from

financial losses by staging an accident, loss of assets or injury to gain financial benefits (Al-Hashedi & Magalingam, 2021). With health insurance, a faulty bill for costly medical procedures can be submitted to the insurer and paperwork for a fake automobile accident can be registered to the car insurance company (Al-Hashedi & Magalingam, 2021). Because of falling prices or natural disasters crop insurance companies face the risk of the customers overestimating their losses (West & Bhattacharya, 2016).

7. Cryptocurrency fraud.

As stated by Al-Hashedi & Magalingam (2021) cryptocurrency is often used by criminals due to lack of regulation and decentralization. This type of fraud deceives people with promises of big profits while providing them with fake investments or services. It mainly targets people with inadequate knowledge about the market and takes advantage of their naivety. These kinds of fraud can generate revenues worth millions of dollars (Al-Hashedi & Magalingam, 2021).

Table 1 presents a brief overview of scientific literature and fraud detection methods used categorized according to types of financial fraud. The fraud detection methods will be presented in more detail in the following subsection.

Table 1. Financial fraud types and related literature with fraud detection methods
(source: compiled by the author)

Type of financial fraud	Literature	Methods
Credit card fraud	Saia and Carta (2019), Madhurya et al. (2022), Shabbir et al. (2022), Chen et al. (2015), Dong et al. (2021) and others.	Fourier transform and Wavelet transform, Support Vector Machine, quantum neural network, RAIN model, ect.
Money laundering	Singh and Best (2019), Fronzetti and Remondi (2017)	Visualization techniques, network analytic techniques
Financial statement fraud	Zhang et al. (2022), Cheng et al. (2021), Shen et al. (2021), Jan (2021)	Support Vector Machine, Random Forest, Naive Bayes, Decision tree, Logistic regression, Recurrent Neural Network, Long short-term memory

Insurance fraud	Bologa et al. (2013), Yan et al. (2020), Amponsah et al. (2022)	Social network analysis, Kernal Ridge Regression, decision tree
-----------------	---	---

1.2. Methods of detecting financial fraud using Big Data Analytics

1. Support Vector Machine (SVM)

Support Vector Machine (SVM) are a supervised kind of Machine Learning technique. A kernel function is utilized to map data to a high dimensional space and to find the hyperplane with the biggest margin among two classes (Madhurya et al., 2022). While using this method for classifying fraudulent credit card transactions it was proven to be efficient, however possible problems arising from unbalanced and differing datasets were mentioned (Madhurya et al., 2022). SVM was not as effective in terms of accuracy as other models examined for identifying loan fraud based on Non-Performing Assets (Attigeri et al., 2021). It was also had the poorest results when performance for imbalanced datasets and specificity were considered (Attigeri et al., 2021). According to Jain et al. (2022) the SVM approach used for identifying credit card fraud did not yield the best accuracy rate results. Shen et al. (2021) compared different models for financial statement fraud detection. Comparatively, SVM technique was not the most accurate when baseline results were considered, although, with inclusion of knowledge graph models it had the highest results for accuracy, AUC as well as F-measure. Zhang et al. (2022) combined SVM and text analytics for detecting fraudulent statements, resulting in a best performing model in terms of accuracy (71 percent). Dong et al. (2018) conducted a similar study with the combination of SVM and text analytic framework to detect corporate fraud. Li (2022) used SVM when analysing e-commerce fraud, concluding that the results of the model, depending on the sample size, were between 60 and 90 percent in terms of accuracy. Dong et al. (2021) utilized SVM based framework for fraud detection in the e-market . When compared to other techniques, SVM based model decreased the error rate for about 20 percent (Dong et al., 2021).

2. Decision tree (DT)

Decision tree - a graphical strategy, with features classified at each node. The main task when building a decision tree is determining the branching criteria (Bi & Liang, 2022). This method is useful when solving complex problems (Cheng et al., 2021). Fraudulent behaviour in the

healthcare industry was analysed by Amponsah et al. (2022) using a decision tree algorithm and blockchain technology. Both pruned and unpruned decision trees were considered. Results showed that the model accurately classified the claims 98 percent of the time (Amponsah et al., 2022). The DT approach was also combined with text analytic framework by Dong et al. (2018) and used to detect corporate fraud. Out of five classifiers considered by Shen et al. (2021) in their paper investigating financial statement fraud, the DT was the most accurate at the baseline. However, a drop in accuracy from around 68 percent to 63 percent was detected when knowledge graph models were introduced (Shen et al., 2021).

3. Random forest (RF)

Random forest is classified as a learning algorithm. Essentially, it is a Bayes classifier, an implementation of the previously mentioned Decision trees (Madhurya et al., 2022). It is composed of many Decision trees, each having a random factor (Cheng et al., 2021). Attigeri et al. (2021) utilized Random forest when analysing loan fraud. This strategy predicted the outcomes as „Performing Asset“ or „Non-performing Asset“. It was deemed to score high with regards to accuracy and a high (over 0.9) F1 Score and Precision index (Attigeri et al., 2021). Zhang et al. (2022) considered the combination of text analytics and Random Forest to test for fraudulent statements, resulting with an accuracy rate of 66 percent. Madhurya et al. (2022) found the Random forest approach to be the most accurate when compared with other five techniques for credit card fraud detection. As a consequence of the imbalanced datasets, sampling algorithms and pre-processing are necessary to classify the data before applying the Random Forest method (Madhurya et al., 2022). To add, Random Forest is not as likely to suffer from noise effects and overfitting as other approaches (Cheng et al., 2021).

4. Logistic model/regression (LR)

The logistic model is meant to analyse the association between free and discrete ward factors (Bi & Liang, 2022). Li (2022) tested the LR approach when dealing with e-commerce fraud and found that, depending on the sample size, the model was accurate between 70 and 90 percent of the time. Tackling financial statement fraud, the logistic regression model proved to be the least accurate (59 percent) out of the five techniques considered (SVM, KNN, DT, NB and LR). The addition of knowledge graph models improved the model's accuracy, though other methods still had higher accuracy ratings (Shen et al., 2021).

Combination of the LR method and text analytics was utilized by Dong et al. (2018) to detect corporate type of fraud. Attigeri et al. (2021) tested the LR model in the context of loan fraud, where, again, it had the lowest (87 percent) accuracy result out of five methods examined (RF, NN, NB, SVM and LR).

5. Neural Network (NN), Recurrent Neural Network (RNN) and Long short-term memory (LSTM) model

The Neural Network technique, composed of many layers, is made up of artificial neurons (Attigeri et al., 2021). Among five techniques (NB, LR, RF, SVM and NN), the aforementioned method was the most accurate and ranked the highest in terms of specificity and precision when used for loan fraud detection (Attigeri et al., 2021). Dong et al. (2018) focused on combining the Neural Network approach with text analytics to test for corporate fraud.

Recurrent Neural Network is a useful tool for examining the relationships between data points (Jan, 2021). Jan (2021) considered the RNN approach and its extension, the Long short-term memory model, as tools for detecting financial statement fraud. The latter model outperforming the RNN one, both in terms of accuracy and precision.

6. Naive Bayes (NB)

Naive Bayes - a predictive model that utilizes prior and likelihood probabilities (Attigeri et al., 2021).

In the context of credit card fraud, this model did not yield the best results in terms of accuracy when other approaches (Fusion model, SVM) were considered (Jain et al., 2022). Attigeri et al. (2021) found that the Naïve Bayes model was the second best, when accuracy and precision were considered, out of five techniques (LR, SVM, NN, RF and NB) tested for loan fraud detection. However, Zhang et al. (2022) found that the combination of text analytics and Naïve Bayes yielded low accuracy (56 percent) comparatively to different approaches (Bag of Words, SVM, RF and NB) when looking to combat statement fraud.

7. Social Network Analysis (SNA)

Social Network Analysis involves gaining additional data on the relationships of a subject. By creating a link between different data sources it allows for predicting fraudulent behaviour as opposed to only detecting it (Madhuri et al., 2021). By detecting a single suspect, a group of fraudsters can be uncovered via the link as this approach takes into account that fraud is usually committed by a group of people, not just a single person (Bologa et al., 2013). By using a network, modelling relationships between major system

information entities, certain indicators are used to isolate suspicious components while performing simulations (Bologa et al., 2013).

8. Fourier transform and Wavelet transform

Saia and Carta (2019) evaluated two novel and proactive strategies used in fraud detection - Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) models. The first strategy, DFT, defines the model in terms of frequency components, while the other (DWT) moves the classification process to a new time-frequency-domain (Saia and Carta, 2019). The considered approaches are shown to perform as good as the Random Forests technique. To add, the aforementioned models are not affected by the scarce and unbalanced data problem, which is why authors suggest testing a hybrid technique making use of the advantages of DFT and DWT methods.

9. Other methods and models

Shabbir et al. (2022) analysed suspicious transactions utilizing quantum neural network, resulting with a high accuracy (97 percent). Artificial Bee Colony algorithm, a method based on Kernel Ridge Regression, was tested by Yan et al. (2020) in the context of automobile insurance fraud and was found to be very accurate (97 percent). Fronzetti and Remondi (2017) examined network analytic techniques (social network metrics) in detecting money laundering activities. Visual techniques were adopted to spot clusters of potential criminals. Money laundering fraud was also considered by Singh and Best (2019) who chose to use visualization techniques to detect unusual bank transactions. Similar approach was utilized by Leite et al. (2020), implementing NEVA (Network Detection with Visual Analytics) to search for patterns of fraudulent behaviour in bank transactions. While Chen et al. (2015) chose to examine transactions with the help of RAIN (Risk of Activity, Identity and Network) model to quantify the risk factor of an object or user.

In conclusion, this section gave an overview of the scientific literature related to financial fraud by describing the main types of financial fraud as well as various techniques of fraud detection. Table 1 highlights the fact that there is little research done in terms of cryptocurrency, securities and commodities and mortgage fraud types. Furthermore, more approaches need to be examined when dealing with different types of fraud. For example, even though Support Vector Machine technique is commonly used, in the literature reviewed, it is only utilized in credit card

and financial statement fraud detection. The main measures used to comparison and evaluation purposes were accuracy, precision and specificity. A common problem of data unavailability was raised by authors, as real world financial data contains sensitive information, it is not commonly accessible.

2. Methodology

In this section describes the statistical methods used in this study to tackle the problem of financial fraud detection. It focuses on finding features that would suggest fraudulent behaviour. These characteristics are highlighted firstly through analysing descriptive statistics, computing and comparing correlation coefficients between a selected characteristic and whether a transaction is fraudulent or not and finally testing whether the relationship suggested by the correlation coefficient is statistically significant.

1. Descriptive statistics – proportions

The first step entails looking at the proportions of fraudulent transactions in different groups expressed in percentages. This percentage will be used to compare the different groups considered and analyse whether there are any significant differences in proportion of fraudulent transactions which could help to identify a characteristic by which fraud could be identified. The different groups considered in this paper were categorized by transaction time(months), gender, transaction place(State). The way of calculating the percentage of fraudulent transactions is as follows:

$$a = \frac{b}{c} \cdot 100 \quad (1)$$

Where b is the amount of fraudulent transactions in a certain group, c is the amount of transactions conducted by the group in total and a is the proportion of fraudulent transactions expressed in percentage terms.

2. Correlation analysis

To further analyze the existence of a linear relationship between a certain characteristic such as gender, place, time of the transaction and fraudulent transactions this study utilizes the correlation coefficient. In this case, Pearson's product moment correlation coefficient was calculated:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2] [n \sum y^2 - (\sum y)^2]}} \quad (2)$$

Where x and y are the variables of interest for which the correlation is calculated – in this case it

will be the binary variable indicating fraud and a certain chosen characteristic. Other variables, n indicates the sample size and r is the correlation coefficient ranging from -1 to 1. The formula calculates the proportion of the covariance between the two variables of interest and the product of the variances of the two variables. Correlation indicates a linear relationship or lack thereof. Important to note that this relationship is not necessarily causal. To establish whether the aforementioned relationship could be causal further statistical analysis should be performed, such as regression analysis.

3. Regression analysis

In order to test whether the relationship between fraudulent transactions and certain attributes is indeed causal and to quantify how strong this relationship is, the following equation was estimated using Ordinary Least Squares:

$$Fraud = \beta_0 + \beta_{1i}x_i + \varepsilon \quad (3)$$

Where Fraud is the binary variable of interest, indicating fraudulent transactions, β_0 is the intercept, x_i is the vector of variables characterizing fraudulent transactions and β_{1i} is the estimated coefficient for each variable, ε is the error term.

By examining the p-values associated with β_{1i} we can determine whether the specific characteristic has a statistically significant relationship with the variable Fraud and to examine the magnitude of this relationship.

fraudulent or not are the transaction amount, the variables specifying the time of the transaction such as the month, day of week and Unix time(in order from the broadest to the most specific) and place of the transaction – the US state and city population. I also take into account the gender of the person performing the transaction and an additional variable indicating the credit card number.

Firstly, the time of the transactions is examined.

Table 2. Distribution of transactions according to months

(source: compiled by the author)

	Fraudulent transactions	Non-fraudulent transactions	Total:
June	133	29 925	30 058
July	321	85 527	85 848
August	415	88 344	88 759
September	340	69 193	69 533
October	384	68 964	69 348
November	294	72 341	72 635
December	258	139 280	139 538

3. Discussion of the Results

The dataset used for executing the analysis consists of simulated credit card transactions performed across USA generated by Sparkov Data Generation. There are both legitimate transactions and fraudulent ones, spanning from the 21st of June, 2020 until 31st of December, 2020. This dataset contains variables indicating the number, date, time, category and amount of the transaction, credit card number, merchant name and location, customer's information(Name, gender, job and date of birth), city population, Unix time and a binary variable indicating whether the transaction is fraudulent or not.

In total there are 555 719 transactions out of which only 2 145 are fraudulent, making up only 0.39% of the whole dataset. This constitutes an unbalanced dataset, as the number of legitimate transactions outweighs that of the fraudulent ones.

Examining the variables given in this dataset, I hypothesize that the variables which could be useful for identifying whether a transaction is

Table 2 present the distribution of the fraudulent and non-fraudulent transactions according to the months. The raw number shows that the most fraudulent transactions were performed in August and the lowest number was in June. In order to be able to compare these numbers it is crucial to take into account the total number of transactions for a particular month. For this reason proportions of fraudulent transactions to the total number of transactions in presented in Figure 1.

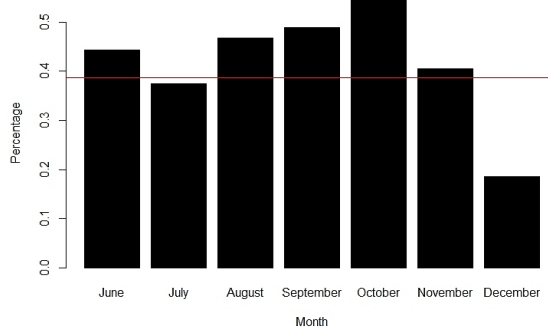


Figure 1. Proportion of fraudulent transactions in each month expressed in percentages (source: compiled by the author)

As Figure 1 presents, the biggest proportion of fraudulent transactions was found in October (0.55 percent), while the lowest was in December (0.18 percent). The low percentage of fraudulent transactions in December can be explained by the large amount of total transactions. Out of the seven months, December had the most total transactions, which can be explained by increased purchases of holiday presents and other necessities. Referring to Table 1, where June had the least amount of fraudulent and total transactions, as only ten days of June were considered, looking at the proportion of fraudulent transactions it is visible that fraudulent transactions made up a much larger (0.44 percent) portion of the total transactions when compared to December.

The red line indicates the average percentage of fraudulent transactions in the whole dataset (0.39 percent). As portrayed by Figure 1, only in the months of July and December the proportion of fraudulent transactions was lower than the average amount for the dataset. An increasing trend is spotted for the proportion of fraudulent transactions from June to October, where after the proportion drops significantly.

Secondly, the proportion of fraudulent transactions is analysed according to the day of the week.

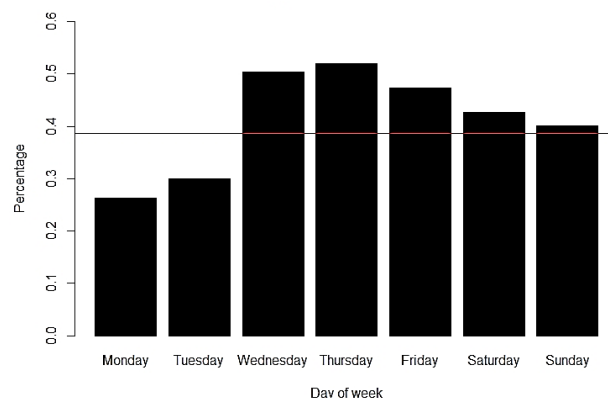


Figure 2. Proportion of fraudulent transactions in each weekday expressed in percentages. (source: compiled by the author)

From Figure 2 we see that Thursday has the biggest proportion of fraudulent transactions (0.52 percent) along with Wednesday (0.5 percent). When compared to the average proportion of fraudulent transactions across the dataset, only Monday and Tuesday fall below this average. Similarly as in Figure 1, we see a raising trend in the proportion of fraudulent transactions from the beginning up until the middle of the week and later a gradual drop. The third aspect which is examined is the place of the transaction. The variable indicating the US state where the transaction took place is used.

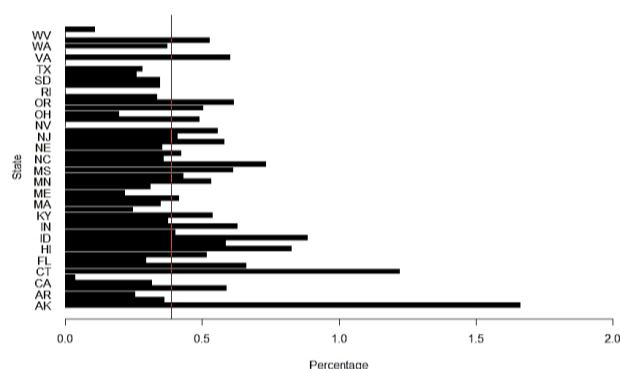


Figure 3. Proportion of fraudulent transactions in each state expressed in percentages (source: compiled by the author)

From Figure 3 we see that the biggest proportion of fraudulent transactions is in Alaska (1.66 percent), Connecticut comes in second with 1.22 percent. Worth noting that in some states (Nevada, Rhode Island, Utah, Vermont and West Virginia) there were no fraudulent transactions at all. When compared to the average proportion of fraudulent transactions across the dataset, there are quite a few states where the proportion is bigger.

The analysis of the proportions of fraudulent transactions by gender follows.

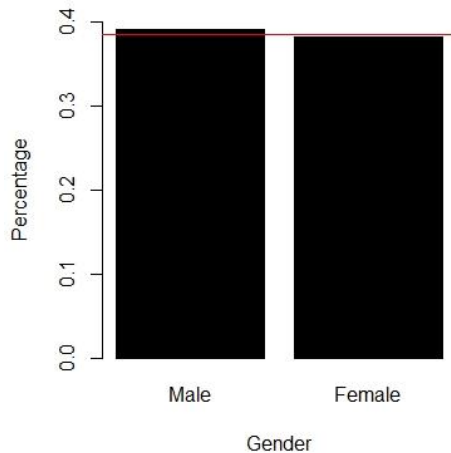


Figure 4. Proportion of fraudulent transactions for each gender, expressed in percentages.
(source: compiled by the author)

Figure 4 suggests that the proportion of fraudulent transactions is slightly higher among male (0.39 percent) than female (0.38 percent). Additionally, the fraudulent proportion for men is above the dataset average, while it is below the average for women.

Additionally, the descriptive statistics for the transaction amount are examined for two types of transactions.

Table 3. Descriptive statistics for transaction amount
(source: compiled by the author)

	Fraudulent transactions	Non-fraudulent transactions
Min.	1,78	1
Max.	1 320,92	22 768,11
Mean	528,36	67,61
St. dev.	392,75	152,47

Table 3 provides the smallest, largest, average transaction amount for both fraudulent and non-fraudulent transactions as well as the standard deviation. The minimal amounts for both types of transactions are quite similar and low, whilst there is a significant gap when comparing the maximum amounts – non-fraudulent transactions include large sums, while the fraudulent ones are kept relatively low. The mean transaction amount, however, is much higher in fraudulent transactions. Considering the fact, that the standard deviation is also much higher in the

case of fraudulent transactions – there are more transaction amounts falling closer to the maximum amount. As we take into consideration that there are much fewer fraudulent transactions and that the amounts of these transactions do not lay close to the mean, when compared with non-fraudulent transactions.

Next, the correlation coefficients between the variable of interest – the binary variable indicating fraud and the numerical variables in the dataset are compared.

Table 4. Correlation coefficients
(source: compiled by the author)

	Fraudulent transaction
Amount	0.182
City pop.	-0.005
Unix time	-0.013
CC number	-0.002
October	0.01
Thursday	0.007
Alaska	0.008
Male	0.001

To explore the relationship between the previously mentioned categories – time, place of the transaction as well as the gender of the person performing the transaction – binary variables were created to indicate transaction made in October, on a Thursday, in Alaska and by a Male. Other variables describing transaction amount (Amount), city population(City pop.), Unix time(Unix time) and Credit Card number(CC number) were taken from the main dataset. The highest correlation can be observed between the transaction amount and fraudulent transaction. The positive correlation coefficient suggests that the variables are moving in the same direction. The correlation coefficient between the variable of interest and both Unix time and October are quite similar in magnitude, although for Unix time the correlation is negative. The lowest correlation is observed for the Male variable. In this case, the correlation coefficients do not suggest strong relationships magnitude wise. Finally, I perform a regression analysis to test the statistical significance of the chosen variables. The third main equation specified in the previous section of

this paper as well as it's modified versions are estimated using Ordinary Least Squares method.

Table 5. Regression results
(source: compiled by the author)

	Dependent variable: Fraudulent transaction		
	(1)	(2)	(3)
Amount	0.0001*** ³ (0.00000)	0.0001*** (0.00000)	0.0001*** (0.00000)
October	0.002*** (0.0002)	0.002*** (0.0002)	0.002*** (0.0002)
Thursday	0.001*** (0.0003)	0.001*** (0.0003)	0.001*** (0.0003)
Alaska	0.012*** (0.002)	0.012*** (0.002)	0.012*** (0.002)
Male	0.00003 (0.0002)		0.00005 (0.0002)
Unix Time	-0.000*** (0.000)	-0.000*** (0.000)	
City pop.	-0.000*** (0.000)		
CC number	-0.000 (0.000)		
Constant	0.229*** (0.022)	0.229*** (0.022)	-0.002*** (0.0001)
Observations	555 719	555 719	555 719
R ²	0.034	0.034	0.033
Adj. R ²	0.034	0.034	0.033
Residual st. error	0.061 (df = 555 710)	0.061 (df = 555 710)	0.061 (df = 555 710)

Table 5 presents the regression results – the coefficients estimated for each variable as standard error given in brackets. For the first model all of the numerical variables that were present in the correlation table (Table 4) were

used. In the second model, five variables with the highest correlation were considered and in the third regression included the four newly created binary variables (October, Thursday, Alaska, Male) with the addition of Amount variable. For the first four variables included in all three models we see no change in the estimated coefficients across different models. All of them are highly statistically significant and the sign of the coefficients are the same in correlation analysis, however when comparing the magnitudes of correlation and regression coefficient the results differ. The biggest coefficient magnitude wise in regression analysis is attributed to the Alaska variable. A much smaller coefficient can be seen estimated for the October and Thursday variables and the smallest one for Amount variable. The variables indicating Unix time and City population are shown to be statistically significant, however, the coefficients estimated for both are close to zero. Other factors, such as Male and Credit Card number were deemed as statistically insignificant.

Considering the models that were estimated, we can see that both R² and adjusted R² statistics are very low, indicating a poor goodness-of-fit of the models considered. It is possible that the aforementioned models can suffer from omitted variable bias when considering the setup of the model as well as a problem such as heteroskedasticity.

Conclusions

In conclusion, this paper presented the problem and relevance of financial fraud, examined and categorized the existing literature on the topic of financial fraud types and detection methods as well as suggested possible factors that can be used to identify fraudulent transactions based on the dataset used and analysed these factors. The conducted literature review presented financial fraud in general and highlighted the main types of financial fraud – bank, insurance and corporate fraud as well as cryptocurrency fraud. After that, the main three types are further categorized into credit card, mortgage, money laundering, financial statements, securities and commodities, insurance and cryptocurrency fraud sections. The overview of past work also considers different ways of detecting financial fraud – Support Vector Machines, Decision tree, Random forest, Logistic model/regression, Neural network and

³ The statistical significance of the estimated is marked via asterisks. Three asterisks meaning statistical significance at 1 percent, two asterisks – at 5 percent and one – at 10 percent.

it's modifications, Social Network Analysis as well as other detection methodologies.

This paper focuses on credit card fraud and uses regression analysis as well as descriptive statistics and correlation analysis. Three models are estimated using Ordinary Least Squares each with different combinations of possible fraud detection factors. Analysing the proportion of fraudulent transactions in the categories of transaction time, place and gender of the person, the binary variables for October, Thursday, Alaska and Male were created as they had the largest proportions of fraudulent activity. The results for all the regression models are very similar – for all the models the coefficients estimated for the variables Amount, October, Thursday and Alaska are all statistically significant at 1 percent, however the magnitude of these coefficients is very small. The largest coefficient can be seen for the variable indicating the transactions made in Alaska (0.012) and although it is larger than the coefficient of correlation between fraudulent transaction and Alaska (0.008), it's impact is small overall. Considering correlation results, the coefficient was able to state the direction of the relationship between the variables, however the magnitudes of the regression and correlation coefficients differs. The strongest correlation observed between the variable of interest and transaction amount was the largest (0.182), however the impact of this relationship has decreased when regression analysis was considered. Both correlation and regression analysis suggest that the relationship between fraudulent transactions and Amount, October, Alaska and Thursday variables exist and the relationships are positive and statistically significant, the impact of these relationships are relatively small.

Considering the results and findings of this paper, further research should expand the dataset and include more variables that can be used to characterize fraudulent transactions – this can improve the overall model as well as minimize the omitted variable bias, problems which are relevant for the models considered in this article. In terms of application, when building a model for detection of fraudulent transactions variables relating to the time and place of the transaction as well as the amount of the transaction are crucial to consider as this paper found a causal association between these criteria and fraudulent transactions. The findings also imply that banks should be very cautious when dealing with transactions from Alaska, especially those made in October and on Thursdays. To test the findings of this paper further, examining a dataset of real-live

transactions would bring valuable insights when compared to the results described in this paper as one the limitations of this study is usage of artificially generated transactions, which may differ from real transactions.

Disclosure statement

The authors the article to do not have any competing financial, professional, or personal interests from other parties.

References

- A. Leite, R., Gschwandtner, T., Miksch, S., Gstrein, E., & Kuntner, J. (2020). NEVA: Visual Analytics to Identify Fraudulent Networks [Article]. *Computer Graphics Forum*, 39(6), 344–359. <https://doi.org/10.1111/cgf.14042>
- Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019 [Article]. *Computer Science Review*, 40, 100402. <https://doi.org/10.1016/j.cosrev.2021.100402>
- Amponsah, A. A., Adekoya, A. F., & Weyori, B. A. (2022). A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology. *Decision Analytics Journal*, 4, 100122. <https://doi.org/10.1016/J.DAJOUR.2022.100122>
- Association of Certified Fraud Examiners (ACFE). (2022). *Occupational Fraud 2022: A Report to the Nations*.
- Bi, W., & Liang, Y. (2022). Risk Assessment of Operator's Big Data Internet of Things Credit Financial Management Based on Machine Learning. <https://doi.org/10.1155/2022/5346995>
- Bologa, A.-R., Bologa, R., & Florea, A. (2013). Big Data and Specific Analysis Methods for Insurance Fraud Detection. [Article]. *Database Systems Journal*, 4(4), 30–40.
- Chen, J., Tao, Y., Wang, H., & Chen, T. (2015). Big data based fraud risk management at Alibaba [Article]. *The Journal of Finance and Data Science*, 1(1), 1–10. <https://doi.org/10.1016/j.jfds.2015.03.001>
- Cheng, C.-H., Kao, Y.-F., & Lin, H.-P. (2021). A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes [Article]. *Applied Soft Computing*, 108, 107487. <https://doi.org/10.1016/j.asoc.2021.107487>
- Domashova, J., & Mikhailina, N. (2021). Usage of machine learning methods for early detection of money laundering schemes. *Procedia Computer Science*, 190, 184–192. <https://doi.org/10.1016/J.PROCS.2021.06.033>

- Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging Financial Social Media Data for Corporate Fraud Detection [Article]. *Journal of Management Information Systems*, 35(2), 461–487. <https://doi.org/10.1080/07421222.2018.1451954>
- Fronzetti Colladon, A., & Remondi, E. (2017). Using social network analysis to prevent money laundering [Article]. *Expert Systems with Applications*, 67, 49–58. <https://doi.org/10.1016/j.eswa.2016.09.029>
- Girija Attigeri, Manohara M M Pai, & Radhika M Pai. (2021). Supervised Models for Loan Fraud Analysis using Big Data Approach [Article]. *Engineering Letters*, 29(4), 1422.
- Jain, H. (2022). Advanced Big Data Analysis Approach for Credit Card Fraud Detection. [Article]. *Grenze International Journal of Engineering & Technology (GIJET)*, 8(2), 1048–1054.
- Jan, C.-L. (2021). Detection of Financial Statement Fraud Using Deep Learning for Sustainable Development of Capital Markets under Information Asymmetry [Article]. *Sustainability (Basel, Switzerland)*, 13(17), 9879. <https://doi.org/10.3390/su13179879>
- Kosemani Temitayo Hafiz. (2016). The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada. [Article]. *CISTI (Iberian Conference on Information Systems & Technologies / Conferência Ibérica de Sistemas e Tecnologias de Informação) Proceedings*, 1, 1090–1097.
- Li, J. (2022). E-Commerce Fraud Detection Model by Computer Artificial Intelligence Data Mining. <https://doi.org/10.1155/2022/8783783>
- Madhurya, M. J., Gururaj, H. L., Soundarya, B. C., Vidyashree, K. P., & Rajendra, A. B. (2022). Exploratory analysis of credit card fraud detection using machine learning techniques. *Global Transitions Proceedings*, 3(1), 31–37. <https://doi.org/10.1016/J.GLTP.2022.04.006>
- PwC. (2022). PwC's Global Economic Crime and Fraud Survey 2022. Protecting the perimeter: The rise of external fraud.
- Saia, R., & Carta, S. (2019). Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks [Article]. *Future Generation Computer Systems*, 93, 18–32. <https://doi.org/10.1016/j.future.2018.10.016>
- Shabbir, A., Shabir, M., Javed, A. R., Chakraborty, C., & Rizwan, M. (2022). Suspicious transaction detection in banking cyber-physical systems [Article]. *Computers & Electrical Engineering*, 97, 107596. <https://doi.org/10.1016/j.compeleceng.2021.107596>
- Shen, Y., Guo, C., Li, H., Chen, J., Guo, Y., & Qiu, X. (2021). Financial Feature Embedding with Knowledge Representation Learning for Financial Statement Fraud Detection. *Procedia Computer Science*, 187, 420–425. <https://doi.org/10.1016/J.PROCS.2021.04.110>
- Singh, K., & Best, P. (2019). Anti-Money Laundering: Using data visualization to identify suspicious activity [Article]. *International Journal of Accounting Information Systems*, 34, 100418. <https://doi.org/10.1016/j.accinf.2019.06.001>
- Sirisha Madhuri, T., Ramesh Babu, E., Uma, B., & Muni Lakshmi, B. (2021). Big-data driven approaches in materials science for real-time detection and prevention of fraud. *Materials Today: Proceedings*. <https://doi.org/10.1016/J.MATPR.2021.04.323>
- West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66. <https://doi.org/10.1016/J.COSE.2015.09.005>
- Yan, C., Li, Y., Liu, W., Li, M., Chen, J., & Wang, L. (2020). An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification [Article]. *Neurocomputing (Amsterdam)*, 393, 115–125. <https://doi.org/10.1016/j.neucom.2017.12.072>
- Yanjiao Dong. (2021). Real-time Fraud Detection in e-Market Using Machine Learning Algorithms. [Article]. *Journal of Multiple-Valued Logic & Soft Computing*, 36(1–3), 191–210.
- Zhang, Y., Hu, A., Wang, J., & Zhang, Y. (2022). Detection of fraud statement based on word vector: Evidence from financial companies in China [Article]. *Finance Research Letters*, 46, 102477. <https://doi.org/10.1016/j.frl.2021.102477>

Submission checklist

Indicate that this submission is ready to be considered by this conference by checking off the following (comments to the director can be added below):

- ☒ The submission has not been previously published, nor is it before another conference for consideration (or an explanation has been provided in Comments to the Director).
- ☒ The submission file is in Microsoft Word document file format.
- ☒ All URL addresses in the text (e.g., <http://pkp.sfu.ca>) are activated and ready to click.
- ☒ The text is single-spaced, uses an 11 point font, employs italics rather than underlining (except with URL addresses), and all illustrations, figures, and tables are placed within the text at the appropriate points rather than at the end.

- ☒ All illustrations, figures, and tables are mentioned in the text and are placed within the text at the appropriate points rather than at the end.
- ☒ The text adheres to the stylistic and bibliographic requirements outlined in this template.

Copyright Notice

Authors who submit to this conference agree to the following terms:

- ☒ Authors retain copyright over their work while allowing the conference to place this unpublished work under a [Creative Commons Attribution License](#), which allows others to freely access, use, and share the work, with an acknowledgement of the work's authorship and its initial presentation at this conference.

- ☒ Authors are able to waive the terms of the CC license and enter into separate, additional contractual arrangements for the non-exclusive distribution and subsequent publication of this work (e.g., publish a revised version in a journal, post it to an institutional repository or publish it in a book), with an acknowledgement of its initial presentation at this conference.

- ☒ In addition, authors are encouraged to post and share their work online (e.g., in institutional repositories or on their website) at any point before and after the conference.

The authors agree to the terms of this Copyright Notice, which will apply to this submission if and when it is published by this conference.