


FACTORS INFLUENCING FRAUDULENT TRANSACTIONS FROM BIG DATA PERSPECTIVE

Fabiana LEVON*, Nijolė MAKNICKIENĖ 

Department of Financial Engineering, Faculty of Business Management, Vilnius Gediminas Technical University, Saulėtekio al. 11, 10223 Vilnius, Lithuania

Received 26 February 2023; accepted 5 April 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 because of fraud, this 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 its 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 detect fraudulent transactions by their characteristics.

This paper is composed of the following parts – firstly, the existing literature of the topic is examined, then the methodology used in this paper is presented. The section discussing the results follows and the paper is finalized by providing conclusions.

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

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 and 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, Al-Hashedi and 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 also 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 and Magalingam (2021) also mention Bitcoin fraud as part of cryptocurrency related fraud.

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

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, etc.
Money laundering	Singh and Best (2019), Fronzetti Colladon 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, Kernel 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 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 used 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 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 outperformed 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 Naive 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 Naive Bayes yielded low accuracy (56 percent) when compared to different approaches (Bag of Words, SVM, RF and NB) while 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 (Sirisha 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 & 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 Colladon 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 for 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

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 total amount of transactions conducted in the group and a is the proportion of fraudulent transactions expressed in percentage terms.

2. Correlation analysis

To further analyse 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. Here, 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.

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 and fraudulent transactions, 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 percent 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, it is hypothesized that the variables which could be useful for identifying whether a transaction is 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 authors)

	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

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.

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.

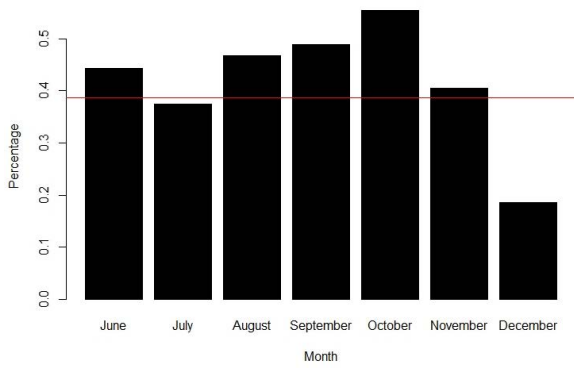


Figure 1. Proportion of fraudulent transactions in each month expressed in percentages (source: compiled by the authors)

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 2, 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 horizontal line indicates the average percentage of fraudulent transactions in the whole dataset (0.39 percent). According to Figure 1, only in the months of July and December the proportion of fraudulent transactions was lower than the average amount for the dataset. The proportion of fraudulent transactions from June to October increases, later the proportion drops significantly.

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

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

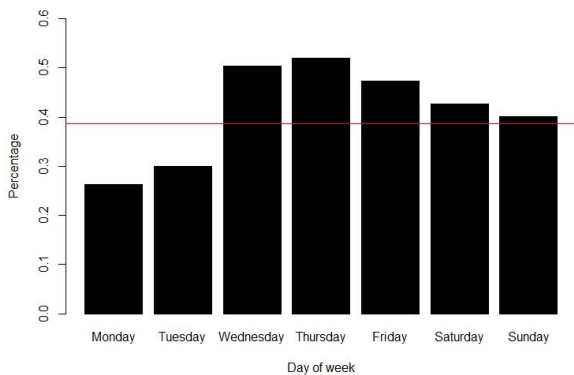


Figure 2. Proportion of fraudulent transactions in each weekday expressed in percentages (source: compiled by the authors)

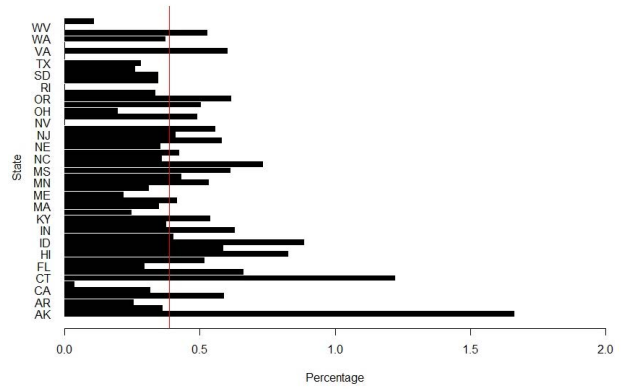


Figure 3. Proportion of fraudulent transactions in each state expressed in percentages (source: compiled by the authors)

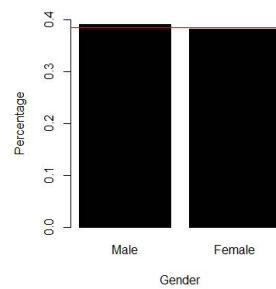


Figure 4. Proportion of fraudulent transactions for each gender, expressed in percentages (source: compiled by the authors)

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.

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 suggests that the proportion of fraudulent transactions is slightly higher among the male (0.39 percent) than female (0.38 percent) gender. 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 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 difference when comparing the maximum amounts – non-fraudulent

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

	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 4. Correlation coefficients (source: compiled by the authors)

	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

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. So, there are fewer fraudulent transactions, but they are of higher transaction amount.

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

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 transactions 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 its modified versions are estimated using Ordinary Least Squares method.

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

	Dependent variable: Fraudulent transaction		
	(1)	(2)	(3)
Amount	0.0001*** ¹ (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 well 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 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 signs 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 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.

¹ 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.

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. 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 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 its 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), its impact is small overall. Considering correlation results, the coefficient was able to state the same the direction of the relationship between the variables as regression analysis, however the magnitudes of the regression and correlation coefficients differs. The strongest correlation was observed between the variable of interest and transaction amount (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, however the impact of these relationships is 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 of the article do not have any competing financial, professional, or personal interests from other parties.

References

- Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *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 health-care 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. (2022). *Occupational fraud 2022: A Report to the nations*.
- Attigeri, G., Pai, M. M. M., & Pai, R. M. (2021). Supervised models for loan fraud analysis using big data approach. *Engineering Letters, 29*(4), 1422–1435.
- Bi, W., & Liang, Y. (2022). Risk assessment of operator's big data Internet of Things credit financial management based on machine learning. *Mobile Information Systems, 2022*, 5346995. <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. *Database Systems Journal, 4*(4), 30–40.
- Chen, J., Tao, Y., Wang, H., & Chen, T. (2015). Big data based fraud risk management at Alibaba. *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. *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. *Journal of Management Information Systems*, 35(2), 461–487. <https://doi.org/10.1080/07421222.2018.1451954>
- 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), 191–210.
- Fronzetti Colladon, A., & Remondi, E. (2017). Using social network analysis to prevent money laundering. *Expert Systems with Applications*, 67, 49–58. <https://doi.org/10.1016/j.eswa.2016.09.029>
- Hafiz, K. T., Aghili, S., & Zavorsky, P. (2016). The use of predictive analytics technology to detect credit card fraud in Canada. In *Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1090–1097). <https://doi.org/10.1109/CISTI.2016.7521522>
- Jain, H. (2022). Advanced big data analysis approach for credit card fraud detection. *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. *Sustainability*, 13(17), 9879. <https://doi.org/10.3390/su13179879>
- 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. *Computational Intelligence and Neuroscience*, 2022, 8783783. <https://doi.org/10.1155/2022/8783783>
- Madhurya, M. J., Gururaj, H. L., Soundarya, B. C., Vidyaashree, 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.* <https://www.pwc.com/gx/en/services/forensics/economic-crime-survey.html>
- Saia, R., & Carta, S. (2019). Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks. *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. *Computers & Electrical Engineering*, 97, 107596. <https://doi.org/10.1016/j.compel-ceng.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. *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. *Neurocomputing*, 393, 115–125. <https://doi.org/10.1016/j.neucom.2017.12.072>
- Zhang, Y., Hu, A., Wang, J., & Zhang, Y. (2022). Detection of fraud statement based on word vector: Evidence from financial companies in China. *Finance Research Letters*, 46, 102477. <https://doi.org/10.1016/j.frl.2021.102477>