Detecting Financial Frauds using Data Mining

Karthik Reddy Gaddam
Dept. of Computer Science
Central Michigan University
Mount Pleasant, USA
gadda1k@cmich.edu

Dayanandh Jangamsrinivas
Dept. of Computer Science
Central Michigan University
Mount Pleasant, USA
janga1d@cmich.edu

Swathi Korrapati
Dept. of Computer Science
Central Michigan University
Mount Pleasant, USA
korra1s@cmich.edu

Chandra Shekar Naik Sabhavat
Dept. of Computer Science
Central Michigan University
Mount Pleasant, USA
sabha1c@cmich.edu

***Abstract*— Based on data mining technology and their use in detecting financial fraud, this paper gives an overview of the research on detecting financial frauds and detecting them. In most cases, data mining techniques such as support vector machines, One-class SVM, bag of words are used to uncover financial statement frauds and insurance frauds. The Boruta algorithm is used for important attribute extraction with the help of random forest classifier. This study provides valuable information for the detection of all types of fraud. Further, the underlying framework offers the capability to solve financial services problem along with privacy and security measures.**

Keywords—Data mining, fraud, credit card, Boruta, insurance, mortgage, statement, support vector machines, bow, word2vec, word embedding.

I. INTRODUCTION

With the advancements in technology, internet has become a mandatory resource for accessing anything remotely. Many businesses have started their own website for ecommerce platforms for making profits using internet. This led to the way of saving time for customers by making payments online. Moreover, there have been rise in number of banking sectors that provide financial needs for customers. All these measures were focused on making the payments electronic. Although convenient, it possesses risk of financial fraud. The internet plays an important role in all these processes. This had led to the growth of financial transactions and have been a target for malicious attacks. Many cyber crimes have started taking place with the invent of electronic payments through these vendors. Financial frauds cause a loss of billions of dollars annually for an organization.

Financial fraud not only affects the businesses but also the customers on regular basis. It is a big threat for many sectors like firms, corporate sectors, and government. It can be defined as an illegal behavior, resulting in a beneficial gain to either individual or organization from unethical or illegal ways. With things like these in mind, many researchers came forth to finding financial fraud detection methods. Although there are many methods for detecting fraud, there's no single method that works best in all the cases. This paved the way to find multiple detection methods that works best for a particular scenario while not good for other scenarios. Any service that includes online transactions is vulnerable to a financial fraud. There, ant-fraud has always been a topic of interest for many researchers.

Data mining is the process of extracting knowledge from a given data. The knowledge thus extracted has patterns that are useful to predict relevant outputs. This is done by using many approaches either single or combined such as statistics, machine learning, artificial intelligence techniques. As far as financial fraud is considered, different kinds of techniques can be applicable, a few being Naïve Bates, Support Vector Machines (SVM), and many more. Besides, outlier analysis does have a crucial role in detecting financial fraud.

II. TYPES OF FINANCIAL FRAUD

Financial frauds are different in many sectors. For instance, frauds can happen in industries, government, and banking sectors. This, paves the way knowing different kinds of frauds. (Hilal & Yawney, 2022)

2.1 Credit card fraud

The concept of performing transaction digitally or electronically without using cash or physical money is called credit. Due to the growth of internet usage, many customers have switched to making payments electronically. Credit card is the most widely used card by the customers. It is a simple plastic card with the electronic chip. This chip is capable of processing the data of the customer. Even though the chip needs power which is not in the card, it is capable of using the power from the machine that performs transactions. Because credit cards are most widely used, fraudsters started targeting customers with credit cards. Credit card fraud can be categorized into: Online and Offline fraud. In Online fraud, the fraudster usually has the credentials of the original card holder stolen. This lets the fraudster do online transactions without a physical card anywhere that has online payments service. In contrast, the Offline fraud involves stealing the credit card from the actual owner and doing the transactions in stores or anywhere with the actual stolen card.

2.2 Mortgage fraud

This fraud usually involves modifying the mortgage documents during the loan application. Usually, a real estate fraud would mislead the value of the property with the intention of gaining value with more material for a less price.

2.3 Financial statement fraud

A financial statement can be defined as an array of records created that explain about the current performance of the organization, including its transactional activities that summarize an overall picture of the organization. These

records are basically used by the banks, stake holders to make decision on approving loans. The main aim of this fraud is to deceive the financial statement users in thinking that the organizations is beneficial and is bankable. Moreover, it also plays a role in increasing the share prices, reducing the tax liabilities, and attract as many investors as possible.

2.4 Insurance fraud

Insurance fraud usually involves deceiving the insurance companies with a false statement. This leads to huge losses for insurance companies while the fraudster who claims the insurance uses illegal methods. Insurance fraud occurs in different kinds of insurance domains such as healthcare, automobile, etc. In automobile insurance claims, the fraudsters usually may provide a fake document containing bills of an accident, which did not occur in the first place. As far as healthcare insurance fraud is considered, the fraudster shows fake reports of medical services and claims the insurance for things like a surgery.

III. METHODOLOGY

This study aims to learn recent data mining applications in detecting fraudulent transactions. We organize the paper by explaining about Bag of words and Word2Vec approaches which are followed by Boruta algorithm for feature extraction. Finally, we explain about Support vector machines and how they are used in detecting anomalies.

3.1 Bag of words

Bag of words is a Natural Language Processing technique which tracks the frequencies of words. Representing the words in numbers is a mandatory task for computers. This is because a computer cannot interpret the words but just numbers. Bag of words is a simple approach that maintains the frequencies of words using data structures like dictionaries. It represents the words in the form of vectors which are of the size of dictionary. Given an unknown sentence or statement, bag of words simply compares the frequencies of two bags, one holding positive, while the other holding negative words. A decision is made based on the frequencies of two bags. It is simple and effective approach in instances like email spam filtering. As mentioned above, initially we have two bags, one holding words of legitimate mails while the other holding the words of spam mails. Given an unknown email, we calculate the frequencies of legitimate and spam words in that mail, sum it up and decide if the mail is spam or not based on the frequencies. That is, the mail is classified as spam, if the word frequencies of spam are higher than the word frequencies of legitimate bag. (Bel, Bracons & Anderberg, 2021)

Although this technique is simple, yet it suffers from sparse matrices. In real scenarios, we encounter different words in emails that are unseen. This is why we end up with sparse matrices. This is why one would consider using bag of words as processing step for representing words in vectors and later use these vectors in machine learning models.

3.2 Word2Vec

Bag of words is a classic model that does not maintain any order of words. Meaning, order is irrelevant. In practical applications, the volume of words tends to be very large resulting in a sparse matrix. This makes many of the vectors meaningless. To overcome this, Word2Vec is used for vectorizing the text. It preserves the positional relationship of words. It is widely used in text classification. Besides, it's based on the word embedding model (Cichosz, 2020). It uses neural networks to learn the word associations from a large corpus of text. Upon learning, it can group the words not just by the synonymous context but also by the semantic context. Word2Vec, as the name implies, represents each word by a vector. These vectors are not just the same as created by the bag of words. They possess the semantic and syntactic qualities of words having their semantic similarities measured by a mathematical function called cosine similarity.

3.3 Boruta algorithm

Feature selection is the process of extracting the set of few effective variables in the large dataset. The dataset contains only a few variables useful for constructing the machine-learning model, while the rest features are irrelevant or redundant. Adding all these redundant and irrelevant variables to the dataset may negatively affect the model's performance and accuracy. Thus, it is essential that the most appropriate features be identified from the data, and that the irrelevant or less important features are removed, which is accomplished through feature selection in machine learning which is critical for fraud detection. An automated or manual approach to selecting the most appropriate and relevant set of features for model building is known as feature selection. This process involves either including or excluding the relevant and important features from the dataset without changing them.

Founded in 2010, the Boruta algorithm is an impressive piece of software that takes a dataset and automatically selects one or more features from it. It was designed as a R package. The Boruta algorithm is also known as a wrapper around the Random Forest classifier (Rung-Ching, 2020). The way it works is it chooses a model that can capture non-linear relationships or interactions like random forest, and fit it on X and Y. Then, the influential variables are extracted from the model by discarding the features that are below the threshold value.

The Boruta feature selection method is based on statistics and performs very well even when no user input is provided. A threshold value is chosen by using two methods: Shadow features and Binomial distribution.

3.3.1 Shadow features

Boruta does not have inter-feature competition. They go up against a random version of them. Take the original data frame features as X, create a new data frame by shuffling the values of the features in the first data frame. Then, the features in the second data frame are called shadow features. Finally, append the shadow data frame to the original data frame to create a new data frame called X_boruta. Then, fit a random forest classifier on X_boruta, Y. Among the original features, the features that have the greatest impact on the model can be selected according to the threshold (Feature Importance). The threshold is defined as the maximum value of shadow feature importance. If the attribute importance is higher than the threshold, then it is

called a ‘hit’. Finally, features are selected when they meet the shadow feature threshold (Kursa & Rudnicki, 2010). The motive is to use the features which can do better than the shadow features.

3.3.2 Binomial distribution

In the above idea, the trial to select useful features is only once. But iteration is the key factor in machine learning. More than one trial may be reliable for selecting the important features. After a successful iteration, we need to decide which feature should be considered. A probability of 50% is the feature's highest level of uncertainty and a binomial distribution is followed by a set of n trials. The binomial distribution between the number of hits and the probability mass function gives three regions: Area rejection region, Area of irresolution, and Area of acceptance. Features that end up in the area rejection region should be discarded because Boruta considers it to be noise. If the features end up in the area of irresolution, Boruta is undecided. Finally, the area of acceptance is where the algorithm considers the features.

3.4 One-class Support Vector Machines

Usually, classification models focus on solving the two or multi-class problems. The goal of machine learning model is to use training data to make best predictions on test or unseen data. However, if we need to train the model on just one class and make predictions on whether the data is similar to training data or not is something that beyond the scope of two or multi-class problems. This is where One class SVM comes into play. One class SVMs are effective against real-time scenarios like analyzing the proper functioning of machinery in a factory. It is not always the case where we can simulate the unexpected problems and give the data to a two-class support vector machine. However, we can always train the model on normal data. Situations like these can be solved using One-class support vector machines. One of the things these models are good at is novelty detection.

3.4.1 The theory of Support vector machines

The SVM classification basically decides the optimal hyperplane among many hyperplanes possible in input or feature space in order to separate positive or negative samples. This decision boundary is built by using maximal margins in input or feature space by using linear separators close to both classes. By learning a decision boundary, an SVM typically minimizes the generalization errors and performs good on unseen tuples. Two categories of data are usually considered when using an SVM: Linearly separable and linearly inseparable. The former case is simple. Here, the data can be separated using a line while the latter cannot be separable. This is where an SVM projects the data into a higher dimension that is practically linearly separable. Once the data is linearly separable, the SVM simply reduces the dimension and resulting dimension has the hyperplane which separates both classes. This projection of the data into higher dimension is done using the kernel trick. The kernel function is designated as:

$$K(x, x_i) = x\,T(x_i) \quad (1)$$

It is not essential to conduct an explicit projection to that space because the decision function just depends on the dot-product of the vectors in the feature space F. It can be substituted with a function K as long as the outcomes are the same. The feature space F can have an indefinite number of dimensions, making the hyperplane used to separate the data exceedingly complex. This is known as the kernel trick and it is what gives SVMs such a huge capability when dealing with non-linear separable data points. Nevertheless, we avoid that complexity in our calculations. Popular choices for the kernel function are linear, polynomial, sigmoidal but mostly the Gaussian Radial Base Function.

3.4.2 One-class SVM using Hyperplanes

The support vector machine for novelty detection usually separates the data points from the origin and tries to maximizes the distance from this hyperplane from the origin. This results in a binary function which is used for classifying the data point based on the density (Qiao, Wu & Peng, 2023).

3.4.3 One-class SVM using Hyperspheres

Another SVM which uses hyperspheres, instead of hyperplanes. This algorithm creates a spherical boundary using the data points. Since the boundary is spherical, it needs a radius $R > 0$. The resulting hypersphere is characterized by the center and radius as distance from center to radius R which also includes vectors on the boundary.

IV. Security and Privacy concerns

Data mining plays a crucial role in extracting useful information from large amounts of data using various techniques. However, the main and initial step for the process is collection of data from various resources (e.g., customers), which may contain sensitive information (e.g., credit card details) and storing them in data warehouses. Now if we consider the question who can access the data security and privacy will become the initial concern in the process.

If we use methods that preserve privacy and extract knowledge from data, then the methods are known as Privacy-Preserving Data Mining (PPDM) techniques. In order to preserve privacy most of the PPDM techniques either remove or modify the original data. PPDM methods maximize the utility of data while maintaining a certain level privacy. Various PPDM methods are used at different levels of process to preserve privacy at each level.

4.1 Data collection privacy

During the collection of data, data privacy is important before we send the data to the collector. So, we use randomization method on each data value we are collecting. Randomization modifies data by adding noise with a known statistical distribution. So, when a datamining algorithm is used the original distribution of data is reconstructed instead of the original data values. Another way to randomize the data is my multiplying the data with some noise of known statistical distribution.

4.2 Data publishing privacy

Few cases exist which are not willing to disclose the ownership and publish the data either publicly or to some third parties for data analysis. In such cases few privacy models like k-anonymity, l-diversity, t-closeness etc., can be applied before publishing the data. PPDM methos use for preserving publishing privacy are known as PPDP (Privacy Preserving Data Publishing).

4.3 Data mining Output privacy

The output of some of the datamining algorithms is very revealing about the input data. So, it is important to preserve output privacy as well. Some of the methods for preserving privacy are: Association Rule Hiding and Downgrading Classifier Effectiveness. While using association rule data mining method, some of the rules reveal about some individual original data. So, by using association rule hiding in such cases is useful in preserving privacy. This method mines all non-sensitive rules making sure that sensitive rules are not discovered. In case of Downgrading classifier effectiveness, there may be leakage of information by some classifier applications. So, to preserve privacy in such cases we downgrade the accuracy of the classifier. Association rule hiding can also be used, to downgrade the classifier performance, as the classifiers sometimes use association rule-based mining as a sub-routine.

V. CONCLUSION

This paper begins with an overview of financial fraud detection and some data mining concepts, followed by a discussion which includes types of fraud detection, their characteristics and techniques used for fraud detection. Fraud detection is the process of monitoring the user's behavior to estimate or detect or avoid unexpected or undesirable behavior. In recent years there has been a significant rise in the number of frauds occurring. The types of financial fraud detection include credit card fraud detection, mortgage fraud, financial statement fraud and insurance fraud. It also explains the characteristics of each fraud type and various techniques that are useful for detection. We also focused on the security and privacy issues that arise while using various data mining techniques. The types of privacy issues and different methods that preserve privacy are also discussed in this paper.

References

Bel, N., Bracons, G., & Anderberg, S. (2021). Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis. *Information, 12*(8), 307.

https://doi.org/10.3390/info12080307

Cichosz, P. (2020). Unsupervised modeling anomaly detection in discussion forums posts using global vectors for text representation. *Natural Language Engineering, 26*(5), 551-578.

https://doi.org/10.1017/S1351324920000066

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. Expert Systems with Applications, 193, 1.

https://doi.org/10.1016/j.eswa.2021.116429

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1–13.

https://doi.org/10.18637/jss.v036.i11

Qiao, Y., Wu, K., & Peng, J. (2023). Efficient Anomaly Detection for High-Dimensional Sensing Data With One-Class Support Vector Machine. *IEEE Transactions on Knowledge and Data Engineering, 35*(1), 404-417. https://doi.org/10.1109/TKDE.2021.3077046

Rung-Ching, C., Dewi, C., Su-Wen, H., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data, 7*(1)

https://doi.org/10.1186/s40537-020-00327-4