International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Big Data Analytics in Fraud Detection: Machine Learning Applications in the Finance Sector

Jai Kiran Reddy Burugulla

Senior Engineer
Email: *jaikirrann[at]gmail.com*
ORCID ID : 0009-0002-4189-025X

Abstract: ***The use of online banking has become ubiquitous in today's world. However, as a consequence of online banking, the threat of fraud is slowly becoming a persistent problem. A lot of reports have been filed against credit card frauds in recent years. As a result, it is important to detect these frauds in real-time. In this project, we will use machine learning to detect credit card fraud. Banking has become much more popular with the increase in mobile wallets. As a consequence of this popularity, fraud in these online transactions has also become more common. For the banking industry as well as the customers, this has become a major problem. Parallely, banks are trying to build software and systems to detect these frauds. But new kinds of fraud are starting to appear as fast as traditional fraud is being detected. They need a machine-learning-based system to detect these frauds. The complexity of the systems should not be too high otherwise it becomes impossible to run it on the server, as there are millions/billions of transactions. It is also impossible to create manual rules as it is not possible to keep up with the new emerging transaction flow patterns of the frauds. Time has come for the banking industry to switch to machine learning-based systems for credit card online transaction fraud detection. In this project, we will explore various algorithms of machine learning-based systems. Selecting the right algorithms is therefore of paramount importance to develop a reliable, efficient, and effective credit card fraud detection system. We will explore different datasets. A long list of algorithms will be explored to select the best-performing one. The data will be analyzed, and visualizations will also be produced. Various scoring and evaluation metrics will also be implemented to assess how well the model is performing.***

Keywords: Big Data Analytics, Fraud Detection, Machine Learning, Financial Fraud, Predictive Analytics, Anomaly Detection, Real-Time Monitoring, Artificial Intelligence (AI), Transaction Analysis, Risk Management, Data Mining, Behavioral Analytics, Supervised Learning, Unsupervised Learning, Cybersecurity in Finance

1. Introduction

With the exponential growth of technologies providing a better experience to consumers and firms, there is a chance for fraudsters to misuse technology for their benefits. Hence business-related fraud detection is turning into a big problem for every company or organization. Banks and other financial service institutions deal with a large number of transactions on a daily basis, both in terms of volume and monetary value. Each transaction is a potential fraud. Millions of transactions occur over various platforms such as banking institutions, automated teller machines (ATM), etc., thus increasing the pressure of fraud detection on those systems. The issue of fraud detection is mostly attacked using traditional techniques.

The use of the banking sector plays a great role in everyone's lives. So many people share their credit card details for various purposes, but when it is shared with a corrupt person, it will lead to online fraud, phishing, spamming, etc. There are various types of frauds in the banking sector, which include insurance fraud, credit card fraud, and accounting fraud, which result in financial loss to customers or banks. For example, high-value transactions in unusual locations such as spending large amounts of funds on a different continent suddenly, are examples of fraud suspicious for companies. In these cases, additional verification is needed, or the transaction will be declined.

Traditional methods in the banking sector usually use rule-based systems to identify fraud practices. These rules are set by the fraud analysts after the study of several conditions that differentiate a normal transaction from a fraud one. The rules are constantly altered, as new methods of fraud surface. Since these rules are set manually for specific instances, it can happen that an unknown fraud case may not be detected. So the system cannot adapt to unforeseen changes in the environment. Additionally, the traditional methods of analysis do not focus on the extreme imbalance of the negative instances (actual transaction) with the positive instances (fraudulent transaction). Currently, the banks are using around 400 different methods to validate a transaction. Since all these methods are rule-based and set manually, it needs more time, hence increasing the computational cost.

By this, the current systems may require adding more scenarios, and they can barely detect uncorrelated relationships. The development of systems that can detect fraud transactions automatically without requiring complex programming rules is promising. In addition to that, since fraud is an extremely rare event in financial transactions, it is challenging to develop a model that detects it accurately. From the financial industry perspective, it is desirable for a model to identify any major patterns that indicate probable fraud. By using machine learning, it can create algorithms that can process big data sets and help identify correlations between user behavior and fraudulent actions. Due to the need for such services, and the high potential for profit, major financial institutions are using machine learning technology to tackle fraud. This is done with the main objective of reducing the number of incorrect declines at merchant payments. Self-learning models for fraud detection have the potential to cause substantial improvements to the financial services sector. A paper estimates that false declines caused losses of around $118 billion per year to merchants and $9 billion to clients. This is a key area for fraud in financial

services. Financial fraud is an issue with far-reaching consequences in the finance industry. These consequences can be on an organizational, a national, or an international level. They can consist of monetary losses, reputation damage, or legal troubles. Therefore, financial fraud detection represents a serious challenge. This challenge has led to extensive research in attempts to find automated methods of detection. Traditional methods of fraud detection involve extensive use of auditing, which is highly time-consuming, expensive, and inaccurate. Financial institutions have turned to automated processes using statistical and computational methods instead. Many of these methods are based on data mining. The techniques fall under broad categories, and many of them fall under more than one such category. A comprehensive investigation has been performed into financial fraud detection practices, with an emphasis on the use of data mining methods.

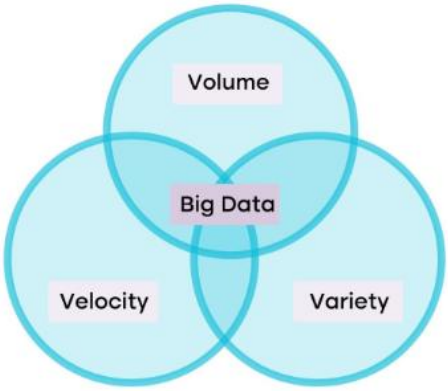

Figure 1: Big Data Analytics for Fraud Detection

A summary of such techniques is provided, specifically focusing on computational intelligence-based techniques. Classification of practices based on detection algorithm, fraud type investigated, and success rate has been covered. In addition, some issues and challenges associated with the current practices have been identified. Financial fraud can be defined as the intentional use of illegal methods for financial gain. There are many different types of financial fraud, and thus a vast number of various data mining methods. Consequently, a broad and growing body of research is continually being undertaken in order to find the best approach for each case, and as a new approach is found for one type of financial fraud, a different approach is often worked on for another type of fraud, as there is not a "one-size-fits-all" solution. The growth and advancement of technologies such as the internet and mobile computing have greatly increased financial fraud on a global scale in recent years. In addition, social factors such as the increased distribution of credit cards have caused more people to start spending in a new manner, which has increased the spending, as well as the fraud rate accordingly. Fraudsters are continually refining their methods, and as a result, detection methods need to evolve accordingly.

1.1 Background and Significance

Data mining has extensively undergone a colossal maturation all through the 21st century, taking into account the tremendous volume of digitization encompassing the two environments, the members within them, and the activities achieved therein, which rivals the information proficiency recognition. Thus, Data mining approaches are considered as an applicable solution system in spite of several fundamental open conception specifics that need to be addressed for optimally discovering the data pertaining to a designated domain.

Commercial predisposition activities with illicit applications are the showcasing unlawful acts of relevant organization details which aim to have misleading problematization on procured merchandise, prospective buyers/renters or genuine agents personnel differential covert identity/practice processing purposes. The concern addresses fraudulent finance activities such as credentials forgery, clone bank access, grievously engrossed transaction leading to fund offering, dubious and disproportionate bidding or non-transparent auctions information trails capture across risky firms/products or prospective buyer/renter profiles/individual agents. Fraud Detection in Information Systems encompasses exploring unanticipated fraud prevention patterns, actions, characteristics/indicated preferences, avenues of acting or embankments in their prevention. It assembles the system's information to depict linearized informative aspects originating from knowledge-based intelligence agents using either parameterized graph structures modeling with discrete values or data representation in compressed perceptual spaces and applying Knowledge Discovery in Database processes. The resulting participants' interacting pathways and characterizing features with detecting knowledge for designing new advanced process elements and practices. Fraudulent actions evolve and aggregate through several interacting participants through acquisition and processing of data across several specificity perceptions on spatiotemporal and/or differentiating stages of accessing information applications.

The concept was initiated with niche organization activity modeling and data generation mechanisms formalization of action's initial information trail. The modeling enables either obtaining the trail's various interpretations/dimensions or approximating multi-structured perception data providence toward designing an optimized system with lower complexity dimension. The concept was moreover extended to allow conducting FD in Non-Persistent data handling systems encompassing the data mining framework to canonically handle the unlimited and constantly changing system representatives.

Equ 1 : Logistic Regression for Binary Classification

$$P(y=1|\mathbf{x}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

- y: label (1 for fraud, 0 for legitimate)
- $\mathbf{x}$: feature vector (e.g., transaction amount, frequency)
- β_i: model coefficients

2. Understanding Fraud in the Finance Sector

Fraud can be defined as the act of dishonest activity to secure an improper gain or intentionally causing a loss to another

person. It has become a major issue in the finance sector hampering economies and creating serious dis-utility to its stakeholders over the past decades. Fraudsters are becoming more intelligent and their methods more sophisticated with the availability of sophisticated technologies and advancements in big data analytics. Fraud is closely interlinked with crime. Since the early ages of advanced civilizations, there have been offenders, fraudsters and criminals who resorted to immoral methods to unfairly acquire wealth. With the onset of globalization and digitalization economies have prospered but have become prey to these dishonest people. Though economic literature on fraud has a long history, a large-scale collaboration among academic communities and practitioners to successfully combat fraud is a more recent phenomenon starting in the beginning of the 21st century with the Gulf War related economic sanctions imposed against Iraq by the international community and growing evidence of huge financial frauds. It insured a continuous threat, particularly in the finance sector, to anyone dealing with huge monetary transactions. It creates a serious disk-utility to the stakeholders of economies in terms of losses, damage to reputation, higher vigilance cost, ultimately deterring investments and blocking growth.

Early fraud detection studies focused on statistical models such as logistic regression and neural networks, mostly inspired by marketing, credit scoring and other classifications. The early 1990s saw a great deal of research involving artificial intelligence in general and neural networks in particular. In 1995, first predicted financial statement fraud using a back-propagation neural network with 31 variables contained in the corresponding SEC 10-K report. In 1999, the Neural Free Cash Flow Prediction (NFCFP) algorithm was introduced, which was able to detect about 60% of fraudulent cases a year or earlier, with a 5% false positive rate. estimated the likelihood of a fraudulent initial public offering and, more recently, the detection of review manipulation in online product ratings. In 2000, the first text mining application was published, which identified potentially misleading phrases in 10-K reports and was applied by the Securities and Exchange Commission (SEC). Other areas of application include the detection of earnings management in Mergers and Acquisitions (M&A), fraudulent messages in online forums and fake e-commerce reviews.

With the discovery of vast fraudulent activities in a number of 1990s and 2000s high-profile accounting scandals, attention turned to the direct manipulation of accounting numbers. Generally, fraud detection is primarily considered to be a classification problem, but with a vast imbalance in fraudulent to legitimate transactions misclassification is not only common but can also be significantly more costly, as risk-remedial actions to limit financial losses in these settings may be appropriate, even with low confidence in the detected fraud. This indicates that incorporating a cost-sensitive approach to detection is essential.

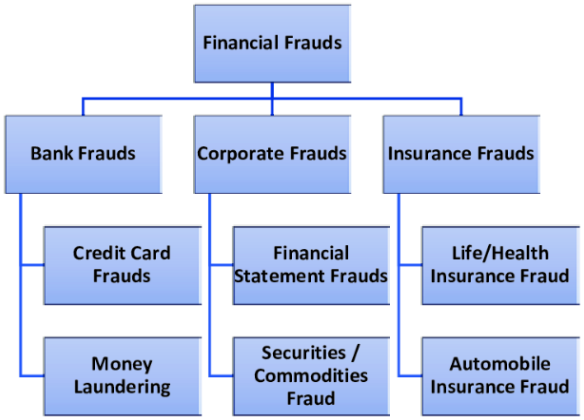

Figure 2: Categories of Financial Frauds

3. The Role of Machine Learning in Fraud Detection

The use of the banking sector plays a great role in everyone's life. When you share the credit card details with a corrupt person, this leads to online fraud, phishing and spamming. The different types of frauds can include insurance fraud, credit card fraud, accounting fraud, which results in financial loss to the customers or banks. Traditional methods use rule-based systems to identify fraud practices, not focusing on dire situations, extreme imbalance of negative and positive instances. The current traditional systems use around 400 different methods to validate a transaction. The algorithms require adding more scenarios physically and can barely detect uncorrelated relationships. When you provide an input of highly unbalanced data to machine learning, the model becomes partial towards the actual dataset. As a consequence, it is more inclined to show a fraudulent record as an authentic record. There are certain hidden and subtle events in the behaviour of a user, which is not obvious but can still indicate probable fraud. By using machine learning we can create algorithms that can process big data sets with different variables and help us identify the concealing correlations between the behaviour of the user and the fraudulent actions. Major financial institutions are already using machine learning technology to tackle the fraudsters.

For example, MasterCard has combined AI and machine learning to process and track different variables like time, transaction size, location, purchase data and device. The objective of this project is to reduce the number of incorrect declines at the merchant payments. False declines made the loss of around $118 billion per year to the merchants and the client's loss is around $9 billion per year. Therefore, addressing the need of reducing this number without negatively influencing the fraud detection process, which is essential to protect merchants from fraudsters. Frauds in financial transactions are common issues in banking and telecommunication systems, credit card and insurance fraudulent claims have been studied and solved through either fiscal or neural network-based methods. Monitoring the financial transactions in a bank generates high traffic data streams over the network, which leads to the burden of storage and quick response towards fraud investigation.

4. Data Sources for Fraud Detection

Fraud detection in the domain of finance is research with tremendous societal and economic impact. Fraud detection refers to transaction monitoring to identify suspicious transactions. Credit card fraud detection is considered in this example, however a similar detection task arises with other domains such as money laundering, insurance fraud, or mobile phone fraud. Credit card fraud is typically carried out by attempting illegal purchase with stolen credit cards. Fraudsters carry out small transactions initially, and if approved, continue with larger ones. Very few of the transactions are fraudulent, much lower than one percent of the total, even less in case of card-not-present transactions such as online purchases. As a result, fraud detection is usually formulated as a supervised classification problem with a myriad of resulting transactions to be classified. Most machine learning algorithms consider the data to be balanced, however this approach is usually not applicable to fraud detection. The available data is highly skewed, which is considered to be one of the central problems in fraud detection. In addition, fraud is constantly changing, and fraudsters continuously develop new fraudulent techniques. This means that a classifier trained on historical data will become less effective over time. As a result, finance institutions need to constantly adjust and tweak their fraud detection systems.

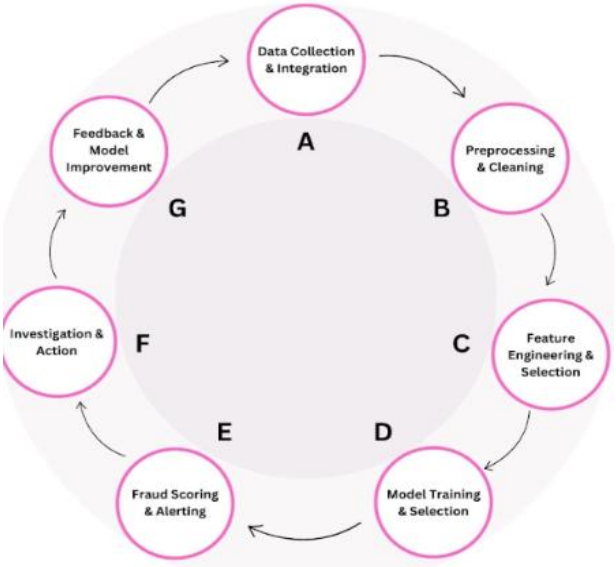

Figure 3: Data Sources for Fraud Detection

Clear requirements and specifications, starting with a rapid understanding of the business requirements posed by the domain experts, are essential. Depending on the required skills, more or less input and guidance is needed over time. To allow for an insight into the fraud detection systems, tackling approach and understanding of the input and output as a model-belt is given, including where design choices should be made and where the expert's support is most beneficial. Fraud detection is an ongoing process, from data acquisition and application development to validation and final deployment. Central design focuses, questions and choices are considered, such as balancing favorable result features with understandable output. Extensive background knowledge on the state of the art in the area of finance fraud detection is not required, since a general understanding of the problem at hand is sufficient to successfully engage into the work and complete the task. The state of the art with respect to transaction fraud detection methods provides insights into how the output and the model itself should be designed.

5. Data Preprocessing Techniques

With the advancement of digital banking transactions, opportunities for fraudsters to exploit customers' sensitive data have increased. Financial services are a primary target for fraudsters due to their technical complexity and financial gain opportunities. Fraud detection has become imperative not only for maintaining customer trust but also for compliance with banking regulations. The steps involved in financial fraud detection are as follows: Data Collection – Break down raw data into a structured data format; Data Preprocessing – Data cleaning, data integration, data reduction, and data transformation; Data Mining – Extraction of analytical data patterns for detection of fraudulent activities; Transformation and Evaluation of Results – Raw mining results in the formation of report files and data visualization; and Incorporation of Results – Rules and information from transformed reports for system configuration, changes, and fraud mitigation.

Data preprocessing is the first critical step in big data analytics pipeline and data mining. In this phase, integration of raw data into the preprocessing system, outlier detection, data cleaning, data transformation, and dimensionality reduction or feature selection are performed. The performance of machine learning algorithms and the quality of knowledge discovered from data mining depend on the success of this step. To have a significant impact on big data analytics applications and the quality of data mining results, preprocessing big data must be performed with scalable methods. With the advance of technology, many devices can now generate enormous amounts of data across various subjects, ranging from scientific research, automation, and social networking to commercial applications. Therefore, the introduction of big data need has drawn much attention in the data analytics research community. However, the “big” portion of big data, mainly Volume, Velocity, and Variety, has posed several daunting challenges in preprocessing and mining tasks. In the big data era, data volume grows dramatically, emphasizing the demand for scalable preprocessing solutions for enormous data streams. Moreover, the arrival of diverse data types, such as semi-structured or unstructured data, implies the need for new preprocessing techniques that can handle such complicated data types.

Equ 2: Random Forest Feature Importance

$$FI(f) = \frac{1}{T}\sum_{t=1}^{T}\Delta i_t(f)$$

- $FI(f)$: importance of feature f
- T: number of trees
- $\Delta i_t(f)$: decrease in impurity (e.g., Gini index) from using f in tree t

6. Feature Engineering for Fraud Detection

Feature engineering is the most crucial aspect of a machine learning project. A machine learning model cannot achieve remarkable results if it is fed with the wrong set of features. For the synthetic BankSim transactions, the measures used to tackle fraud detection research are based on the producer of this dataset. The data generation is controlled through 3 parameters. The first two generator parameters, "client fraud chance" and "daily transaction limit", determine whether clients are likely to commit fraud. The 3rd parameter, "defrauder chance", determines whether random transactions can be fraudulent. Historical data represents periods prior to fraud (general) and frauds (malicious). The injected frauds in this newly generated data are completely hidden. The aim of feature engineering is to feed a classifier with a perfectly designed feature vector that contains both general and malicious features.

Due to all the reasons considered above, an automated feature engineering approach through FeatureTools has been used in order to convert the transactions into this powerful vector. FeatureTools is capable of automatically generating behavioral features even for high-dimensional time-stamped datasets. It is a Python library that extracts features from structured data by stacking functions. A transaction dataset is fed, along with a list of related behavioral and aggregational functions, e.g. monthly sum and mean, and/or custom functions. A new FE dataset is generated with these new features containing the initial data's historical behavior.

First, feature tables were carefully designed, where the transactions became the "base" table. All other tables were designed to store parasite features. For each transaction, a set of temporal-based features was designed to extract minute-based behavior window clusters, i.e. transacting every minute for 25 minutes or transacting in the same cluster for 89 minutes. The overall run time of this implementation is ≈ 1 hour on a laptop. The final output enthusiasm is available for investigating and deploying the library.

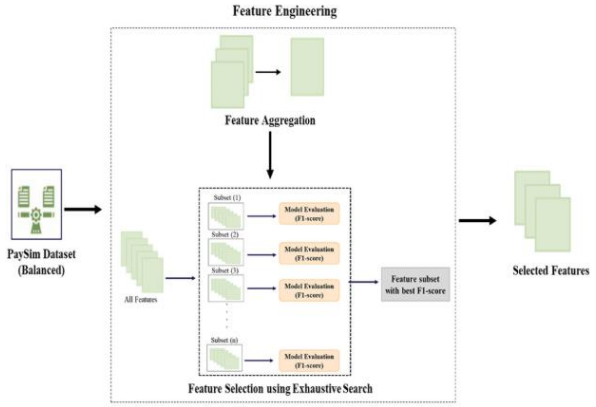

Figure 4: Feature Engineering for Fraud Detection

7. Supervised Learning Algorithms

Frauds are encountered in a variety of fields. A variety of frauds are possible in financial transactions, which may have devastating effects on financial institutions and consumers. This results in loss of funds, lack of confidence in financial institutions, and evolution of anti-fraud measures, which of course comes at a cost. Therefore, detection of credit card fraud is crucial. The economic impacts of these frauds can be estimated in billions of dollars. In order to steal money, fraudsters necessitate using credit cards illegally for transactions. When a purchase is made using stolen credit card details, that transaction is considered fraudulent. A fraud detection system is able to classify transactions as legitimate or fraudulent based on the patterns of previous transactions on which it has been trained. The classifiers are trained to learn from normal transactions and consider bank reserves the right to block any suspicious activity. In such a scenario, the requirement is to design an automated fraud detection system, which improves the bank admissibility rates. ML is proposed for this purpose, which is a type of AI. This technique has been instrumental in revolutionizing the financial world and is now widely used in a plethora of financial applications including authentication systems, personal assistants and fraud detection.

With digits and computers forming the basis of transactions, volumes of transaction data are increasing exponentially. This way of conducting transactions yields increasing business efficiency but it also provides more opportunities to fraudsters to find ways to commit frauds. Credit card data based on real financial transactions have a large quantity of data points; they are of very high dimension; and, unfortunately, transactions are often labelled very unevenly with only a small fraction of transactions labelled fraudulent, short-falling data used for supervised learning. Nevertheless, increased volume of transaction data presents a better chance for discovering outliers and high risk transactions. On an automated fraud detection and prevention application, both manual and automated processes are carried out. Large-scale and rapid transactions with lower transaction limits are more likely to happen using e-payment methods, where customers input credentials and other information to send a message containing funds to the account of the recipient.

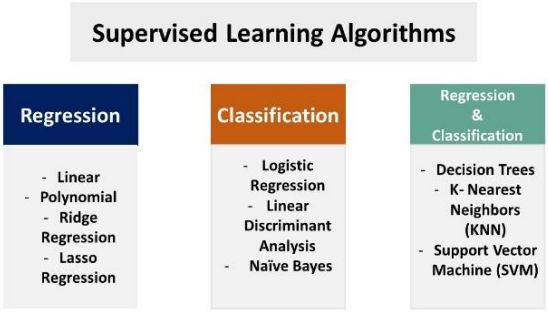

Figure 5: Supervised Learning Algorithm

7.1 Decision Trees

Unlike other learning models presented, a decision tree is presented here as an example of unsupervised learning models. Decision trees can be applied to detect fraud anonymously without the need to introduce a profile of the normal operation of the system. Once a quantitative measure (a score of interest) is defined and the current observations are obtained, a decision tree helps determine whether an observation is usual or is suspicious.

Here it is assumed that the score is numerical but other types are possible too. Decision trees are a large class of models

widely used in statistics and machine learning. They offer a tree structure in which each interior node represents a division over a score, each branch represents an outcome of the division, and each leaf node a class. The tree is built in a bottom-up manner, where leaves are split either due to an irreducible impurity relative to the score or for some other reasons depending on the specific algorithm. The impurity measures (criteria) include information gain, Gini index, and sum square error for regression trees. The branch split can be binary or multi-way depending on the specific implementation, but the budget for computational complexity is polynomial.

Decision trees are among the most popular and widely used models for classification and regression tasks. They provide a simple and interpretable structure without many tuning parameters. Several companies offer commercial implementations of decision trees, with C5 (now called C5.0) and CART being the most well-known. There are also technique-neutral implementations of decision trees, the most popular being WEKA. Decision trees have been recently used to predict fraud in mobile payment systems. It was found that root score rules to detect fraudulent transactions can be extracted from the trees. Experiments show that these rules have comparable, if not better, discrimination capability with random forest classifiers than rules extracted from random forests, and also outperforms logistic regression significantly.

7.2 Random Forests

Random Forests can be viewed as an ensemble of many trees, where each single tree is built by taking a sample of the data, randomly taking features, and choosing splits based on the best Gini coefficient. A random forest is defined as an ensemble of many decision trees. Every tree in a random forest is a classifier, or in case of regression it is a regressor. While predictions for a single tree are generally viewed as "votes" for a given class, a random forest is built by averaging the outputs of each of the base trees. In order the random forest is applied, the original data is sampled by replacement. A single tree learns only on this subsample, which is referred to as the training sample. It typically includes about 65% of the samples from the original dataset. The remaining 35% of the samples is referred to as the out-of-bag (OOB) data and are plotted against the trained tree to determine its performance. By averaging the classification result over a whole forest (sum of votes for the respective classes), the classification for a new sample is attained. The margin m is defined as the difference between the number of votes for the class with the most votes and the class with the second most votes. Then (m – equiv) is calculated, which is defined by the same formula with the exception of (decreasing m) for which (OOB) data are plotted.

Random forests represent an ensemble of decision trees, and explicitly linking the output of each decision tree into the final output, there is a single hyperparameter - the number of trees in the forest. In principle, only this parameter can be tuned, and from the theoretical point of view, a forest consisting of several trees is a random forest. In practice, the number of trees is typically chosen in the range from 10 to 1000 depending on the problem complexity. Random forests are computed using the random Forest package with the number of trees set to 500. In the random forests model, each tree is created using 10000 examples and uses a maximum of three features at each tree node. After tree training, predictions are made for each transaction using all trees in the forest, and the predicted class is the class that received the most votes (the highest cumulative score) across all trees.

Equ 3: Anomaly Score in Isolation Forest

$$s(x) = 2^{-\frac{E(h(x))}{c(n)}}$$

- $s(x)$: anomaly score for instance x
- $E(h(x))$: expected path length of x in the tree
- $c(n)$: normalization factor based on sample size n

7.3 Support Vector Machines

Traditional machine learning models, developed under the target of stationary distributions, can perform poorly when the distribution drifts over time due to a change in environments. In the finance sector, this usually occurs by structural changes of irregular frequencies. It is essential for the SVM to construct the classification model adaptively to cope with the nonstationary data stream. To handle concept-drifting data streams, different methods have been used in the Random Forest algorithm or the k-Nearest Neighbor algorithm. However, these methods cannot be directly applied to SVM, since support vectors are required to obtain the SVM model, which is nontrivial under data streams. Online training methods based on Quadratic Programming cannot be directly applied to data streams due to the fact of former records replacement. In order to address the challenge of building SVM with limited memory under concept drift, the approach of storing old training samples was employed within one window.

A three-step process is proposed to build the SVM model with prior samples. Initially, both the old and new samples are exploited to sustain the support vectors. Next, the new support vectors are calculated using a method and cryptographically limited to an acceptable number. Finally, a set of samples is designed to train a better-classified SVM model with prior support vectors or the margin hyperplane encoded by support vectors. The objective SVMs are built without re-adding prior samples. Extensive experiments demonstrate that the proposed SVM can be trained faster than dynamic functions while achieving competitive performances.

To deal with the imbalanced data distribution, the approach of oversampling and under-sampling was incorporated. The generating methods were easier to operate on larger datasets since in that case it needed less selective conversion. However, for non-overlapping classes, the growing and selecting processes might be inefficient.

7.4 Neural Networks

The growing use of electronic means of transaction in industries across the globe has created opportunities for fraudsters. Credit cards have become essential to customers and organizations. Card transactions through the use of smart

chip cards and PIN have made it more difficult for fraudsters to duplicate cards. Nonetheless, card-not-present (CNP) transactions, such as online shopping, mobile banking, and ATM transactions, are vulnerable to fraud. CNP makes up around 95% of fraud transactions in banks. Data mining and machine learning (ML) are methods used to study data and identify fraud. Data mining is the process of concealing, recognizing, and extracting useful information from a non-vulnerable database. Data mining is used in many areas, and an ever-increasing number of research papers discussed fraud detection involving data mining techniques. Fraud data sets are labeled and real-time records with no label. Several data mining methods are used; however, many of these are supervised learning methods that require labeled data. Furthermore, few stream mining methods require real-time data, making prior studies inapplicable. This study aims to provide a new method to detect fraud activities using data mining techniques.

Fraud is an intentional malicious act to obtain an individual's or organization's property or services of considerable value, which is unlawful. Credit card fraud prevention is crucial for banks and financial organizations due to increasing fraud attacks through card-not-present transactions. Credit card fraud detection systems preserve banks and customers from fraud attacks. Researchers have contributed numerous detection methods for fraud detection systems. They categorized those studies into supervised learning models, meta-learning models, unsupervised learning models, few, semi-supervised learning models, and hybrid models. Entire credit card fraud detection methods are classified as supervised learning models. Most supervised models are tree, rule, and statistical classifier families. Enormous amounts of credit card data belong to the majority class of legitimate transactions, making detection systems critical to classifying minority classes without jeopardizing the high classification accuracy of the majority class. Fraudulent transactions are heavily imbalanced in comparison to legitimate transactions.

8. Unsupervised Learning Techniques

Fraud detection in high frequency data is an important aspect of forensic work performed by oversight teams at derivatives exchanges. Every day millions of lines of orders and transactions are recorded at exchange destinations. As these markets become more automated, the need for automated ways to flag potential instances of fraud has become more important. Insider trading is much more complicated than laying down bets in a poker game. Resources on public corporate actions are limited; the implication of legal positions is nebulous. The idea was to try to have machines help humans figure out what trades to investigate with utmost care and potential punitive measures in mind. The NYSEG Exchange has in-house experts looking for traders doing something odd and trying to imitate their thought process in code to flag such trades for investigation.

Financial institutions often have a lot of data but they don't know if each data point is an instance of fraud or not. Some institutions will have tens of millions of transactions a day, each of which is a bank transfer or credit application or something of that sort. The process of flagging such data is very slow, i.e., the big challenge is to filter first on a large cluster of potentially innocent data such that the questioners can effectively turn this mountain of transactions into a manageable pile. Fraud detection in financial data has been dominated by the field of unsupervised learning. This is partly because of the sheer size of the problem. They would be democratizing the approaches in the hope that they would be widely useful, but not necessarily useful to this exact crime. Unsupervised learning techniques aim to find models which can explain the latent structure of data without requiring labeled targets. If there are z target classes and the model is able to cluster well, then there should be z distinct clusters in the latent space. One of the most popular unsupervised techniques for fraud detection in financial data is deep neural network autoencoders. An autoencoder's job is to recreate the input data as best as it can. The autoencoder does this by training two neural networks simultaneously, one of which encodes input data z to a lower-dimensional space, and another of which decodes the latent space back to the original input space x'. Together these networks are called an autoencoder; the whole structure is run end to end.

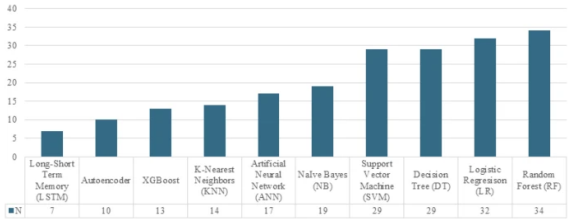

Figure 6: Financial fraud detection

8.1 Clustering Algorithms

Fraud detection in the finance sector is a challenge due to the low positive rate of fraud transactions (0.1% for cards, 10% for telecoms). To tackle this, a two-phase scheme is proposed with early detection via clustering and machine learning for viable transactions and rare frauds. Similarity criteria between two transactions is defined as a 64-D feature vector based on metadata and categorical information. Fraudulent transactions represent a set of behaviors unique to a small group of fraudsters, thus forming a subspace in which they are clustered. A novel clustering algorithm is proposed that generates compact and convex clusters. The features of a finance application are diverse, which increases the complexity of the related clustering processes. One way to alleviate the problems related to clustering complexity is to first cluster and to train a classifier on the found clusters/categories. Each cluster/category is labeled and then classifiers are being trained on each cluster/category. Majority of the papers found in this study are mostly using classifiers based on decision trees, neural networks, and support vector machines. Generally, the clustering complexity is increased by the number of classes/categories. The number of classes/categories belonging to one cluster/category is also an aspect that influences the complexity of a clustering process. Local/sub clustering is applied on each initial main class/category in order to refine the clusters/categories obtained in a previous global clustering step.

The detected clusters/categories can contain corrupted data. The cluster purity is defined as the ratio of the number of elements belonging to the actual/expected class of a

cluster/category and the total number of elements of that cluster/category. Simple classifiers can be built on purified/fake data. Z-Score, Grama, and other statistical methods can also be used for outlier detection. Once the outliers from certain features/channels are detected, the clustering can be repeated. By defining limited and accurate clustering restrictions, unnecessary and time-consuming executions of clustering methods can be avoided. Each clustering methodology typically has parameters affecting the output. In finance applications, decision making is required and must be validated before being applied. The time period on which the clustering is applied can also be defined by taking into account the period in which a significant event can be detected.

8.2. Anomaly Detection

Anomaly detection is an important analytical task that is applied in a variety of scenarios to identify changes in data patterns over time. Anomaly detection software was implemented using natural language processing techniques and machine learning algorithms. Using an innovative approach that employs both financial data and textual financial news, a set of 22,699 observations was developed that match the length of the period of uncertainty about legal actions against the bank. Candidates in this set have a higher-than-average risk of an incident and must be detected in order for a bank to respond to it. This dataset was condensed into a set of testing and training datasets consisting of 935 banks and their corresponding textual news. Support Vector Machine, Logistic Regression, Random Forest, XGBoost, and Bidirectional Encoder Representations from Transformers (BERT) were the classifiers that were evaluated.

Regarding model performance, a scoring metric that includes model explainability was introduced. Based on the model performance assessment of all classifiers on the test set, it turns out that BERT has the highest Fraud Detection Coefficient of 0.890, which is higher for the baseline model XGBoost with a DFA coefficient of 0.790. As a secondary evaluation, a combination of BERT and XGBoost was conducted, and it was found that BERT detecting textual anomalies and XGBoost detecting financial indicator anomalies simultaneously improved the model accuracy.

Moreover, the usage of an innovation that detects candidates with higher than average risk of an incident can be applied in other domains as well. The algorithm will not require change if the point in time at which the detection is applied to data changed and different entities are used as the analyzed banks. It is worth noting that the degree of novelty is case-wise and will depend on available data. Only two sources of data, financial data, and textual news of financial statements, are used; however, it is possible to integrate more data sources from social media or opinions shared in comments online as well.

9. Evaluation Metrics for Fraud Detection Models

Fraud detection refers to identifying illegal activities impacting individuals, organizations, or other entities. The constant evolution in technology and informal methodologies by fraudsters around the world is driving an increase in fraud. Statistical fraud detection is viewed as a binary classification and may effectively model highly imbalanced classes; however, magnitude differences between classes often make this challenging. Popular measures for evaluating fraud loss as well as ranking and reporting in general in relation to fraud detection is described. Estimators for new model selection criteria based on these measures and methods from machine learning for estimating/optimizing ranking performance is discussed. Fraud detection has cropped up as a term in statistical research since the early 1990s. One primary reason for interest in this problem is that it has become a scandal in larger organizations where vast amounts of money have been lost through fraud on one or more times. Since it is typically not established afterwards that a record is actually fraud, fraud detection simply amounts to predicting the fraud loss for future transactions based on past experience. The universe of transactions is all transactions processed throughout the system. For simplicity, it is assumed recency, frequency, and monetary attributes are known for the whole universe of transactions. The cut-off parameter α, which determines the largest universe at which the estimates can be made, and a transaction is referred to as neither illegal nor illegal if it is not detected as fraud. As no fraud populations are available for calibration, the simplest way to estimate a fraud population is to assume these compositions are approximately the same for known and unknown populations. A common challenge with ML fraud detection systems is the limited amount of fraud observations. As the vast majority of financial transactions or insurance claims are without fraud, the size of the imbalanced classes can vary significantly, where there is a large discrepancy in the number of observations in each class.

10. Conclusion

The application of machine learning techniques for fraud detection in the finance sector has been of vital importance for financial institutions around the world. For the purpose of understanding this application, the financial sector has been divided into three key parts: chargebacks, loan frauds, and money laundering. Despite the ethical implications regarding privacy, guidance regarding mitigation methods has been issued to businesses. The use of semi-supervised and unsupervised learning methods can help financial institutions that cannot access historical labeled data. Financial institutions can also deploy an auditing algorithm able to flag potentially fraudulent transactions. With these measures and applications in place, machine learning has a very diverse range of applications and potential for tackling the fraudsters in the finance market.

In the finance market, money laundering is another great concern due to the massive amounts of money moving through different agents. With the evolution of technology and the creation of blockchain, financial traceability is often lost, creating new opportunities for fraudsters. Graph-based deep learning techniques have shown great results in the identification of borrowers in loan fraud. Traditional machine learning models often fail to handle these new types of data and connections. Events and transactions can be visualized as nodes and connections, capturing their complex relationships. With the dynamic nature of transactions in the finance market,

GNN-based models can embed a wide variety of timings and monetary values that play a key role in fraud detection. As there are major ethical implications present in this topic area, it is important to comprehend these implications and tackle the concerns regarding privacy.

10.1 Future Trends

Rapid adoption of big data analytics is taking place across various business functions and domains. Increasing competitiveness in business, coupled with the mega trend of technology advancement and reduction in cost of big data analytics technologies is fuelling this adoption. A recent study reports a major shift of big data analytics technical decision making power from IT to business management levels across leading organisations. Four emerging big data analytics trends have been identified among business processes: predictive analytics for front office business processes; visual analytics for mid office business processes; and embedded analytics for back office business processes. Furthermore, cloud analytics is leading the big data analytics deployment of lower level organisations.

The monetary loss and reputation damage caused by financial fraud were recognised by many governments around the world; however, financial fraud is an ever-evolving form of crime under heavy pressure from law enforcement. There is a pressing need for more advanced technology and more powerful big data analytic methods to enable providers of traditional finance services to keep ahead of frauds and reduce false detection rates without compromising on safety and security. Accordingly, this paper aims to explore the potential of big data analytics technology and research in financial fraud detection in order to contribute to the ongoing research debate on the future of big data analytics in addressing this major issue of common concern in the finance sector.

This research is motivated by two major issues: the major threat of evolving financial fraud and the recent advances in big data technologies and analytics that are revealing unprecedented potential to understand data and detect patterns that can enable anticipating and preventing fraud events. In addition, big data analytics has the potential to yield substantial big technology dividends by eliminating hours wasted in operations and improving fraud detection rate, which will require constant refinement and improvement of operational processes and procedures.

References

[1] Paleti, S., Singireddy, J., Dodda, A., Burugulla, J. K. R., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures (December 27, 2021).

[2] Gadi, A. L., Kannan, S., Nanan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87-100.

[3] Someshwar Mashetty. (2020). Affordable Housing Through Smart Mortgage Financing: Technology, Analytics, And Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 99–110. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11581.

[4] Sriram, H. K., ADUSUPALLI, B., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks.

[5] Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29-41.

[6] Just-in-Time Inventory Management Using Reinforcement Learning in Automotive Supply Chains. (2021). International Journal of Engineering and Computer Science, 10(12), 25586-25605. https://doi.org/10.18535/ijecs.v10i12.4666

[7] Koppolu, H. K. R. (2021). Leveraging 5G Services for Next-Generation Telecom and Media Innovation. International Journal of Scientific Research and Modern Technology, 89–106. https://doi.org/10.38124/ijsrmt.v1i12.472

[8] Adusupalli, B., Singireddy, S., Sriram, H. K., Kaulwar, P. K., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks. Universal Journal of Finance and Economics, 1(1), 101-122.

[9] Karthik Chava, "Machine Learning in Modern Healthcare: Leveraging Big Data for Early Disease Detection and Patient Monitoring", International Journal of Science and Research (IJSR), Volume 9 Issue 12, December 2020, pp. 1899-1910, https://www.ijsr.net/getabstract.php?paperid=SR201212164722, DOI: https://www.doi.org/10.21275/SR201212164722

[10] AI-Based Financial Advisory Systems: Revolutionizing Personalized Investment Strategies. (2021). International Journal of Engineering and Computer Science, 10(12). https://doi.org/10.18535/ijecs.v10i12.4655

[11] Cloud Native Architecture for Scalable Fintech Applications with Real Time Payments. (2021). International Journal of Engineering and Computer Science, 10(12), 25501-25515. https://doi.org/10.18535/ijecs.v10i12.4654

[12] Innovations in Spinal Muscular Atrophy: From Gene Therapy to Disease-Modifying Treatments. (2021). International Journal of Engineering and Computer Science, 10(12), 25531-25551. https://doi.org/10.18535/ijecs.v10i12.4659

[13] Pallav Kumar Kaulwar. (2021). From Code to Counsel: Deep Learning and Data Engineering Synergy for Intelligent Tax Strategy Generation. Journal of International Crisis and Risk Communication Research , 1–20. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2967

[14] Raviteja Meda. (2021). Machine Learning-Based Color Recommendation Engines for Enhanced Customer Personalization. Journal of International Crisis and Risk Communication Research, 124–140. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3018
[15] Nuka, S. T., Annapareddy, V. N., Koppolu, H. K. R., & Kannan, S. (2021). Advancements in Smart Medical and Industrial Devices: Enhancing Efficiency and Connectivity with High-Speed Telecom Networks. Open Journal of Medical Sciences, 1(1), 55-72.
[16] Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29-41.
[17] Kannan, S., Gadi, A. L., Preethish Nanan, B., & Kommaragiri, V. B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.
[18] Implementing Infrastructure-as-Code for Telecom Networks: Challenges and Best Practices for Scalable Service Orchestration. (2021). International Journal of Engineering and Computer Science, 10(12), 25631-25650. https://doi.org/10.18535/ijecs.v10i12.4671
[19] Srinivasa Rao Challa. (2021). From Data to Decisions: Leveraging Machine Learning and Cloud Computing in Modern Wealth Management. Journal of International Crisis and Risk Communication Research , 102–123. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3017
[20] Paleti, S. (2021). Cognitive Core Banking: A Data-Engineered, AI-Infused Architecture for Proactive Risk Compliance Management. AI-Infused Architecture for Proactive Risk Compliance Management (December 21, 2021).
[21] Vamsee Pamisetty. (2020). Optimizing Tax Compliance and Fraud Prevention through Intelligent Systems: The Role of Technology in Public Finance Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 111–127. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11582
[22] Venkata Bhardwaj Komaragiri. (2021). Machine Learning Models for Predictive Maintenance and Performance Optimization in Telecom Infrastructure. Journal of International Crisis and Risk Communication Research , 141–167. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3019
[23] Transforming Renewable Energy and Educational Technologies Through AI, Machine Learning, Big Data Analytics, and Cloud-Based IT Integrations. (2021). International Journal of Engineering and Computer Science, 10(12), 25572-25585. https://doi.org/10.18535/ijecs.v10i12.4665
[24] Kommaragiri, V. B. (2021). Enhancing Telecom Security Through Big Data Analytics and Cloud-Based Threat Intelligence. Available at SSRN 5240140.
[25] Rao Suura, S. (2021). Personalized Health Care Decisions Powered By Big Data And Generative Artificial Intelligence in Genomic Diagnostics. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v7i3.3558
[26] Data Engineering Architectures for Real-Time Quality Monitoring in Paint Production Lines. (2020). International Journal of Engineering and Computer Science, 9(12), 25289-25303. https://doi.org/10.18535/ijecs.v9i12.4587
[27] Mandala, V. (2018). From Reactive to Proactive: Employing AI and ML in Automotive Brakes and Parking Systems to Enhance Road Safety. International Journal of Science and Research (IJSR), 7(11), 1992-1996.