IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

Big Data and Machine Learning in Fraud Detection for Public Sector Financial Systems

Dwaraka Nath Kummari[1], Srinivasa Rao Challa[2]

Software Engineer, dwarakanathkummari@gmail.com, ORCID ID: 0009-0000-4113-2569[1]

Product Technical Lead, srinivas.r.challa.sm@gmail.com, ORCID ID : 0009-0008-4328-250X[2]

Abstract: Fraud detection remains an active area for research. While fraud is a difficult problem hampered by issues ranging from uncertain data to adversarial environments, new technologies and techniques from the fields of data science, machine learning, and big data bring opportunities to alleviate some of the difficulties. In this paper, fraud detection is examined in the context of financial systems for public-sector applications. Public sector financial systems face challenges with fraud detection, such as a large volume of transactions and a need for complex monitoring rules based on the context of transactions. The analysis of public sector financial systems is framed as a data-driven approach to understand the domain. Initial steps to facilitate the analysis of data within a public-sector context are taken by proposing a model framework consisting of detectors, monitors, and pattern mining techniques along with an input data requirements and output results taxonomy. The frameworks' components are investigated and elaborated on in terms of the public sector financial systems. Furthermore, future directions toward further development and evaluation of the components within the context of public sector financial systems. Fraud detection remains an active area for research. While fraud is a difficult problem hampered by issues ranging from uncertain data to adversarial environments, new technologies and techniques from the fields of data science, machine learning, and big data bring opportunities to alleviate some of the difficulties. In this paper, fraud detection is examined in the context of financial systems for public sector applications. Public sector financial systems face challenges with fraud detection, such as a large volume of transactions and a need for complex monitoring rules based on the context of transactions. The analysis of public sector financial systems is framed as a data-driven approach to understand the domain. Initial steps to facilitate the analysis of data within a public sector context are taken by proposing a model framework consisting of detectors, monitors, and pattern mining techniques, along with an input data requirements and output results taxonomy. The frameworks' components are investigated and elaborated on in terms of public sector financial systems.

Keywords: Anomaly Detection,Predictive Analytics,Behavioral Modeling,Supervised Learning,Unsupervised Learning,Real-time Monitoring,Data Integration,Risk Scoring,Entity Resolution,Natural Language Processing (NLP),Data Lake Architecture,Feature Engineering,Graph Analytics,Model Explainability (XAI),Regulatory Compliance Analytics.

I. INTRODUCTION

Fraud is the act in which a person intends to obtain a benefit, which may involve damage to another party. Both public and private organizations have to take measures to prevent fraud, which increases its complexity and costs. Organizations need to actively search for fraud and take preventive measures, which weighs heavily on the available resources. Hence, important first aspects to investigate are the specific costs of fraud for a certain organization and what resources are available to mitigate this fraud. Better insights into the tradeoff between fraud costs and detection efforts could stimulate the development of transparency mechanisms.

To this end, the public sector can be distinguished into two sub-domains. The first being governmental organizations that have a clearer picture of their fraud risk and a dedicated department with analysts and auditors. And the second being public services who may have a less clear picture regarding what types of fraud are present and/or may lack dedicated staff to mitigate it. Nevertheless, here too personalization and fraud is an increasing cost. Therefore, it is interesting to investigate when detection is successful and what suppliers exist to alleviate the burden on small organizations which may not have the knowledge, expertise and resources to act against fraud. The above is all highly relevant from a practice point of view; however, scientific literature concerning fraud in public sector financial systems organizational fraud detection is limited in quantity and many papers also do not distinguish between public and private sector in its measures and typology. Since better consideration of this domain can lead to new avenues for both practitioners and researchers, this gap in literature serves as the basis for the recent research.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

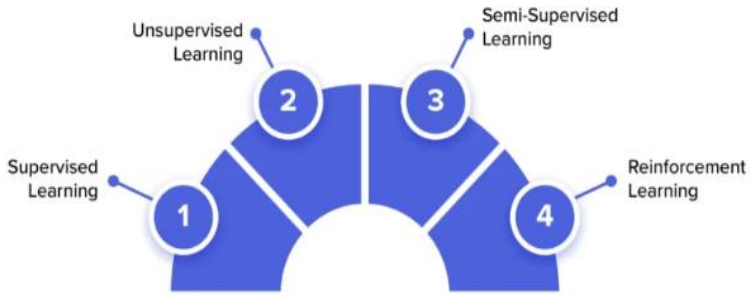

Fig 1: An Analysis on Financial Fraud Detection Using Machine Learning

1.1. Background And Significance

Integrity is crucial in any citizen-centered financial systems, including Budget Service, that handle transactions involving public fund. Unfortunately, fraud attempts and schemes exist and thrive that can bring down agencies and policies of the nation. Hence, real-time fraud detection in financial systems has gained more significance for public sectors. Fraud detection plays a critical role in the overall trust of society towards agencies of the nation. With the increased use of accounting and banking services in multiple aspects, there is a growing number of erroneous transactions and abuse of online transaction services to conduct fraudulent activities. Financial fraud detection refers to the use of techniques and systems to detect and prevent fraud. With information technologies advancing rapidly, financial fraud also becomes more diversified and complicated, leading to a disaster of the economic system and reputation of enterprises and society of nations. As a major sector of the economy, government and other accounting-paying platforms, non-involvement of fraud can not only save a huge amount of money, but also increase citizen's trust.

Equ : 1 Fraud Probability via Logistic Regression

$$P(\text{Fraud}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

- **Purpose:** Estimates the probability of a transaction being fraudulent.
- **Variables:**
  $X_1, X_2, \ldots, X_n$: Features such as transaction amount, vendor ID, time of day, etc.
  $\beta_0, \beta_1, \ldots, \beta_n$: Model coefficients learned from data.

II. UNDERSTANDING FRAUD IN PUBLIC SECTOR FINANCIAL SYSTEMS

Financial fraud has been a widespread and growing problem for many years. Recent arrests of major financial executives have brought financial fraud to the forefront of regulatory and media attention. However, the sophistication of such fraud schemes rises as the industry and regulatory agencies progress. As a result, effective fraud detection methods are needed to prevent and mitigate such undetected financial fraud schemes, given the sheer volume of transactions processed across financial databases. Financial fraud can involve criminal acts such as conspiracy, insider trading, half-truths, and market manipulation. White-collar offenses involve forgery, theft, fraud, or conspiracy to defraud by bank officers and employees, accountants, and other persons. Fraud detection may not be concerned with legal prosecution; however, legally material frauds are of greatest concern to forensic accountants.

The detection schemes used to prevent and investigate frauds can fall into two categories: traditional schemes (indicators, benchmarks, monitoring, and statistical tools) and intelligent schemes (artificial intelligence tools, expert tools, neural networks and related tools, and decision support systems). The statistics around financial fraud summarize the concerns. There have been over 140 arrests in the US due to financial crimes since the beginning of the global credit crisis. The SEC has begun investigating the records of the CEOs, CFOs, and other high-ranking employees of multiple companies due to suspected insider trading. Regulatory agencies have "frozen" portions of major investment banks and funds due to suspicions of Ponzi Schemes, resulting in substantial public outcry.

These events have raised questions about the effectiveness of existing fraud prevention and investigation policies at multiple levels. While these traditional indicators of fraud have served a purpose, their effectiveness in preventing monetary losses from future frauds is questionable, as evidenced by the recent concern over financial fraud. There is a critical need for novel detection methods, capable of identifying even the most elusive frauds prior to investigation or prosecution.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

2.1. Types of Fraud

Broadly, fraud is defined as the tendency to abuse a set of resources for personal advantage or gain. The type of fraud perpetrated is often dependent on the means and systems in place within organizations and is often a subset of the following general types:

1. Misappropriation of Assets: The theft of any organization's resources is typically termed asset misappropriation. Such fraud is frequently accomplished by a single employee and in cases of large fraud involving a larger set of operators, such as dishonesty in newspapers, collusion with a vendor is often involved. In this situation, the vendor pays certain bribes to some stakeholders within the organization in exchange for the fraudulent activity, which allows them to take advantage of the organization's assets.

2. Corruption: In the cases of corrupt activities, means of cheating are created to undermine the organization's objectives. Under such schemes, personnel can charge lower amounts of products bought as per their needs by inflating the actual value of goods supplied and accepting bribes from vendors. It is important to note that the purchased goods may well be unused goods or sham items, which can result in substantial losses to organizations.

3. Financial Statement Fraud: Organizational performance is largely reflected in a set of statements, termed financial statements. Misrepresentation of one or more counts within such statements to create a false image of performance or net worth would fall under such a category of fraud. Such fraud typically requires collusion within a relatively larger group. For unlisted organizations, materiality thresholds are often waived in fraud scenarios, resulting in death spirals of exposure in fraudulent activity and bankruptcy.

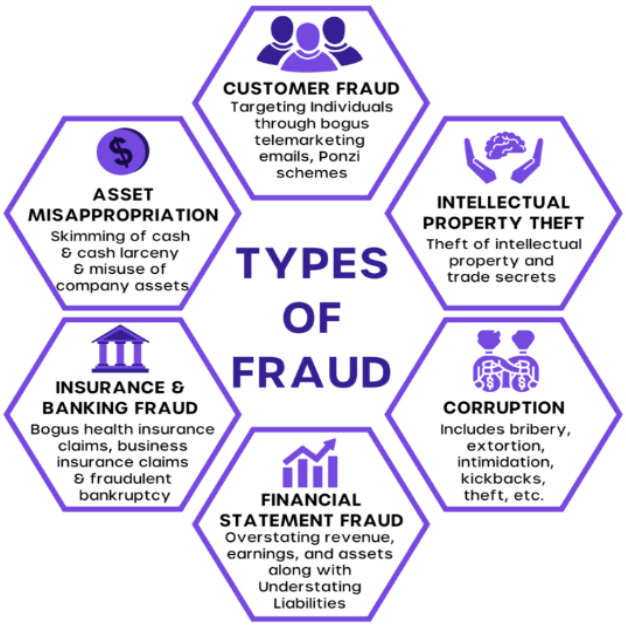

Fig 2: Fraud detection and prevention

2.2. Impact of Fraud on Public Trust Increasing public awareness of corruption scandals has dramatically influenced the public perception of International Financial Institutions (IFIs). In this perception, there are two interconnected concepts of trust and distrust, which can ultimately determine the legitimacy and credibility of these financial institutions. Due to their longstanding existence and prominence, IFIs wield substantial influence over government manipulation, acceptability in the eyes of the international community, commodity prices, and the overall financial integrity of a country. Corruption scandals surrounding IFIs will ultimately result in reduced external trust and deteriorated cash inflows from workplace organizations and fund donors. Reports of corruption on the international front can lead to a complete collapse of public trust. Financial institutions are responsible for public money, as they play a crucial role in channeling tax revenue to the State Treasury, making pension payments and social welfare spendings, and borrowing money. All these procedures require the utmost trust from the general public, as confidentiality and privacy are transferred from the private sector to public finance. For public trust to develop, it is necessary for public servants and organizations to be perceived as fair, fostering a belief in non malfeasance. Then, the need for trust arises, and as public trust develops,

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

lower public participation and accountability empowers governments to practice discretion, which enables opportunities for rent-seeking and embezzlement in organizations. In public trust-distrust systems, accountability and public participation play a crucial role in ameliorating corrupt governments. Therefore, the negative feedback concept has an impact; as awareness of corruption scandals increases, public distrust increases.

Financial institutions are responsible for public money since they are heavily involved in the channelling of tax revenue to the State Treasury, the making of pension payments and social welfare spendings, and borrowing money from the money market. Public trust in cooperation with the financial sector guarantees continuity and success in depriving criminal activity of its financial basis. However, money laundering processes are highly lucrative and in part suitable for large organizations at all levels. At the international level, corrupt countries lose investment opportunities and other beneficial relationships, which indirectly benefit European countries, too. Overall, it is a fourfold problem since loss of money means loss of purchasing power, investment and stimulation of the economy, but also the loss of trust. Without trust a country cannot establish pricing mechanisms, peace, and stability, and wide-scale distrustful administrative cultures will characteristically be corrupt.

III. THE ROLE OF BIG DATA IN FRAUD DETECTION

Financial fraud detection is a pressing issue in finance that can have far reaching consequences, including loss of reputation, stockholder confidence, bankruptcy, class action lawsuits, and loss of client funds which leads to emotional stress. The traditional methods of detection are extensive use of auditing, which is time consuming since the dataset is extremely large and hence inefficient, and only prior knowledge can be applied for formulating heuristics during the manual auditing process. Additionally, the current manual auditing methods that are employed in the banks do not replace any control measures in the system. Fraud detection is a very hard issue since the data being analyzed is typically very large in volume, and the fact that frauds happen now and then. With the rapid being taken up by the financial institutions of automated processes utilizing statistical and computational methods, an urgent need arises for an overview and survey on the different methodologies being put to use in the domain of fraud detection for a better understanding of the issues involved. During the last three years or so, banks, insurance companies, and credit providers have set up large automated systems for the detection of possible frauds within the financial and insurance industry. As an input to this system, user behavior, transaction data, caller number, and other available auxiliary data are collected by log files which contain information of activity in anomaly detection. Then on the basis of predefined rules, preprocessing and clustering, the detected anomalies generate alarms that come in the view of a fraud analyst. These analysts examine there alarms for further discrimination on case level with the help of specialized user interfaces.

Financial fraud can be defined in any number of ways, but the intent behind one specific, legal definition is crucial for the understanding of what fraud is. Financial fraud can also be defined as the intentional use of the variable or illegal methods for the purpose of obtaining financial gains. Financial fraud can be classified into many forms, i.e., performance fraud, fabrication fraud, misrepresentation fraud and abuses of discretion. There exists a number of types of financial fraud or misconduct and a huge variety of data mining methods capable of detecting or preventing them. The financial fraud moves from the simple, conspicuous, institutional fraud that can be caught through sheer human scrutiny of the transactions, to the complicated misconducts that are more subtle and covert. All of these demands novel intelligent fraud detection approaches. With the rise of the Internet and mobile computing, which enhance the capability of real-time transactions, there has been an explosion in the volume of transaction data generated by such businesses in banks, credit card companies and insurance companies. This leads to more fraud in financial domains and thus more need for automatic intelligent detection.

3.1. Data Sources and Types

As government and public sector financial systems increasingly rely on automated environments to process large volumes of transactions, they stand to benefit from better, data-driven technology for fraud detection. However, such technology requires consideration of the scope and nature of the analytical tasks performed, as well as the underlying data, data flows and data management to support them. Participants from large Australian public sector agencies took part in workshops that uncovered their fraud detection needs and practices. Their data needs, in particular, were expansive and multifaceted, touching on systems, schemes, actors, transactions and features, and requiring a wide range of heterogeneous data sources. This section reports on the data sources, types, flows and management as prerequisites for better fraud detection. Financial systems in the public sector are transactional systems. These are the systems run by agencies to process data on business activities. Each transaction usually has its own database row containing information such as identity attributes for actors, description attributes for schemes, etc. In this database context, data can be regarded as a set of tables, with columns as record features and rows as records to be processed. With better analytical capability, the fraud detection

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

need has evolved to both improve existing, and develop new, detection capabilities. Such fraud detection capabilities can be categorized into methods, and applications. In general, methods are complex by themselves and are mostly suspected to have better identification accuracy through enhanced analysis algorithms. On the other hand, methods can produce resulting duplicate transactions, which can be classified to make them more meaningful for users. These two aspects are referred to as the insight discovery and knowledge transmission of data analysis, respectively. For tracking and stopping fraud attempts, users need to extract data from the regression environment, feed it into a library tracker tool for updating definitions/rules, and redistribute the enhanced rules to all grading systems.

3.2. Data Volume and Velocity The advances in technology have generated a plethora of data. In finance infrastructure systems, the data is collected at the transactions level such as transaction amounts, fiscal year, payer ID, host EIN and SSN, recipient ID, host transaction type, description, etc. The volume of the transaction data is huge. For example, the financial data of a state in the U.S over 10-20 years is larger than 100 GB. The data is not only large but streamed real-time, which adds difficulties and challenges to its analytical tasks. The state financial data is generated and recorded in real-time, and new data is streamed, written, and appended. Thus, the data dimension, data volume, and data velocity make the analysis systems complicated and challenging. As data is huge, instead of searching through the whole data, data summarization is employed to enhance efficiency.

The data is also streamed in real-time. Real-time online calculations should be adopted in the analytical systems to support the continuous assessments. The efficiencies of the analytical systems need to be improved. Currently, the financial systems of public sectors and municipalities in most states in the U.S. are based on off-the-shelf products, which are not designed for in-depth exploration or data analysis. Therefore, the analytical tasks are performed with limitations on large systems of off-the-shelf products, resulting in inefficiency and inconvenience. New analytical systems need to be built that are able to efficiently and effectively conduct the analysis in the coming age of big data and machine learning/CAD frameworks.

An overall architecture of the new analytical systems is designed. The analytical functions include particular analysis to facilitate the auditing, and exploratory analysis to find anomalies, frauds, or outliers. As data is recorded in real-time, the schemes are only designed for online performance. Employing efficient statistical measures or pre-built models avoids calculations on the full data. Fast data streaming techniques such as sketch methods and histogram statistics are devised to facilitate real-time calculations. Integrate services enhancing the deployment and normalization for big data visualization techniques helps user-friendly exploratory analysis assisted by expert recommendations. The performance of different approaches is extensively compared.

Equ : 2 Risk Scoring Function for Transactions

$$\text{RiskScore}(t) = \sum_{i=1}^{m} w_i \cdot f_i(t)$$

- **Purpose:** Combines multiple fraud indicators into a single score.
- **Variables:**
  $f_i(t)$: Fraud indicator function for transaction t (e.g., frequent use of same vendor).
  w_i: Weight for the indicator (from ML model or expert tuning).

IV. MACHINE LEARNING TECHNIQUES FOR FRAUD DETECTION

Machine learning (ML) techniques became an important technology adopted by leading companies and institutions in various sectors, such as finance, banking, and social networks, to detect fraud related to new customer behaviors, commands, or transactions. Different algorithms have been employed to train a model and classify those transactions as either fraudulent or not, but the majority of studies in recent years employ a small number of ML classification algorithms. Hence, it is important to evaluate the performance of a variety of classification algorithms to eliminate fraud in transactions.

Prior to discussing these ML methods for fraud detection, the fundamentals of each method must be described. Logistic regression is a family of statistical methods to analyze a dataset in which there are one or more independent variables that determine an outcome. Commonly used with binary outcomes, the response is usually denoted as “success” with value 1 and “failure” with value 0. Support vector machine classification has proved to be efficient for multiclass problems and for that reason has become popular. It generates a hyperplane that divides the two data classes using maximal margin, and the decision function is produced by a linear combination of support vectors, which are a subset of training samples around the separating hyperplane. The K-nearest neighbor classifier assigns a test sample to the class dominated by the K closest training samples, which are called its neighbors.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

Training of K-NN is easier, as it only learns the distance between training samples; however, since the learning is lazily done, its classification can be time consuming when the dataset is large. Furthermore, the class is decided according to the closest K neighbors, but there can be problems with misclassification when they belong to different classes. Enhanced versions of this algorithm such as weight-dependent K-NN can deal with size-imbalance problems to improve performance.

Decision trees recursively perform a feature split until reaching terminal nodes which contain only data in the same class. The decision about how to split the feature proportionally favors a variable that can minimize the impurity measure of the split. This method is interpretable, inexpensive, and makes no assumptions about data distribution; however, it is sensitive to noise and can be prone to overfitting, especially for small imbalanced datasets.

A random forest uses bagging of bootstrapped samples to build decision trees where every time a new tree is constructed, a different set of features is randomly selected without replacement and only those features are tested. Forests improve generalization and reduce overfitting by averaging across a large number of trees; however, they require careful attention to parameters that are model dependent.

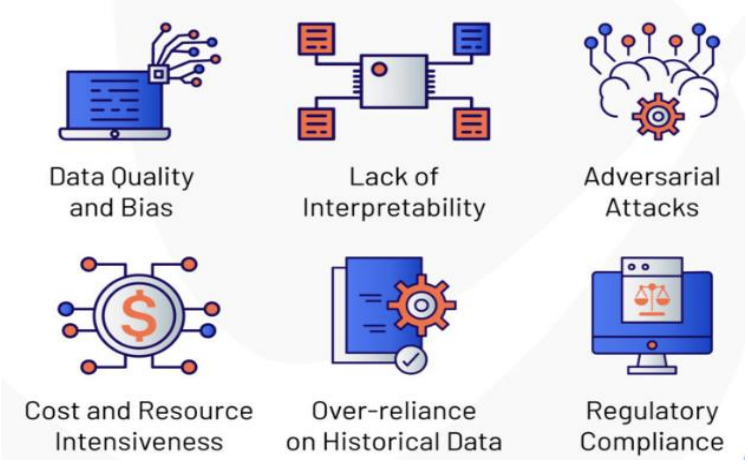

Fig 3: Fraud Detection An Overview of Techniques

4.1. Supervised Learning Approaches In most supervised approaches, anomalies, non-anomalies, or both are pre-identified using expert knowledge. Selected features and labeled data are used to train models to detect anomalies that have not been previously detected. This includes the use of classifier algorithms such as Naïve Bayes, Decision Trees, and support vector machines (SVMs). These algorithms, however, do have some drawbacks. Beforehand knowledge of anomalies is needed. It is often the case that the amount of labeled data is drastically smaller for the positive class than for the negative class, which is often referred to as the class imbalance problem. As a result, there is a need to develop models that use text mining techniques to generate labeled data.

In supervised approaches the class of the observations is identified. Once the data is labeled, supervised classification algorithms can be employed. Labeled observations can be used to train probabilistic or deterministic classifiers. Supervised anomaly detection search for regions in the feature space in which few observations occur relative to the rest of the observations. Binary classifiers that aim to detect outliers from the non-outliers can be used as well. Data is partitioned so that observations in the same partition (class) are similar while observations belonging to different partitions are different. Examples are k-Nearest Neighbors, Support Vector Machines, and certain clustering algorithms.

4.2. Unsupervised Learning Techniques

Fraud detection is an active research topic where a proper method can greatly reduce the loss of a financial institution. High dimension financial data give rise to challenging detection tasks, where anomaly detection is commonly done in unsupervised learning. Therefore, fraud detection in public sector financial transactions is an interesting domain to study anomaly detection with various machine learning techniques.

Verification of large-scale transactions in public sector finances has recently become a new research topic in financial fraud detection. Public sector transactions have strict rules and properties where unusual transactions can be suspected

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

of being fraudulent acts. Massive amounts of resultant monitoring data accumulate systematically in public procurement systems and thus the need for developing a fraud detection system arises. Although such acts are punishable by law it is mostly the responsibility of the expertise of civil servants to discover them rather than an automatic system. The legislation and rules of transactions may differ from country to country, making the problem unique, where rules of the system in transaction processing and monitoring are very complicated.

Thus it is important to take a special focus on such rules or properties in a model, aiming for both event detection and explanation of findings. The fraud acts in public procurement transaction data should generate unusual submission patterns in bids, continuously long run absence from bids in contractors, or recurrently high run-up connected bids which are common in public sector tenders. Conversely models in unsupervised learning which are trained solely based on normal data distribution give rise to better generalization ability when modeled properly.
The goal of fraud detection in financial data has been mainly dominated by unsupervised learning approaches. Unsupervised learning techniques aim to find a model that can learn and explain the latent structure of the data without requiring labeled targets, making them the best candidates for deploying on financial transaction data in fraud detection.

4.3. Deep Learning Applications

The amount of cybercrimes like various financial crimes has risen significantly due to digitalization. Fraud detection has attracted major research interest as a recent popular research topic in Machine Learning (ML) and Artificial Intelligence (AI). A case study on fraud detection in credit transactions is provided as background for information-analyzing and predictive-decisioning tasks on automatic fraud detection in non-proprietary transaction networks. Various recent techniques in statistical and machine learning methods have been reviewed, and ways to evaluate and benchmark these techniques via suitable performance metrics have been suggested. Effective statistical and machine learning methods have played a significant role in fraud detection for identifying various fraudulent behaviors including e-commerce credit card fraud and telecommunication fraud. These technologies would show their adaptability and effectiveness in the evolving context of the fraudulent activity across various financial industries for fraud detection and periodically updating the rules of fraudulent transactions. These techniques would continuously play a crucial role in protecting the society from serious cybercrimes like financial crime and security breach. A framework based on Financial Technology with Artificial Intelligence and Blockchain has been proposed in an effort to study the principles on which electronic banking fraud detection measures can be taken.

Global financial crime activity is driving demand for machine learning solutions in fraud prevention. Collaborative learning advances, which offer the potential for improved performance in payment fraud prevention systems whilst avoiding data sharing between disparate organizations in the financial ecosystem, are rare. A collaborative deep learning framework for fraud prevention, designed from a privacy standpoint, is presented. In collaboration with a consortium of four major financial organizations, there is a wide variety of financial crime signals, data, and collaborations that facilitate NGOs to combat financial crime. Security risks and privacy issues related to sensitive financial data are identified, and a vision for secure federated learning across a consortium of financial organizations is presented. A collaborative deep learning framework, designed from a privacy standpoint, is outlined. Latent embedded representations of varied-length transaction sequences, along with local differential privacy, are leveraged in order to construct a data release mechanism which can securely inform externally hosted fraud and anomaly detection models. The effectiveness of the contribution is assessed across two distributed data sets donated by large payment networks, and robustness to popular inference-time attacks is demonstrated.

V. DATA PREPROCESSING FOR EFFECTIVE ANALYSIS

With increasing complexities in financial ecosystems, the government-private sector is leaving no stone unturned to safeguard financial fraud detection solutions in various areas such as Money Laundering, Remittance Fraud Detection, Accounting Fraud Detection, Data Integrity, Contract Requirements, High-risk Payments and much more. In some cases, this has resulted in various tools being acquired/outsourced that may not have inbuilt consideration to optimise all such areas. This paper presents Automated visualisation and clustering tools that can aid in investigation of suspected financial fraud areas offering numerical and visual evidence that can complement the overall solution.

The fact that the fraud detection community is generally sceptical about tightly ‘plug-n-play’ automated solutions is well understood. As a consequence, the majority of current solutions are limited to a specific area or require a big amount of training, monitoring, feature engineering, intelligence and orchestration effort to be effective. There is a strong intent from both the government and sector to future ‘automate’ or develop further use cases on existing solutions. For either of those cases, it is of paramount importance to highlight beyond-the-market-wide factors that make it difficult that

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

existing solutions keep delivering expected financial fraud detection results. Outsourced solutions are often limited to a functionality that had been anticipated early and their vendor's vision and offered ecosystem are not taken into consideration when embedding it in operations. As a result, integrative analytics and visualisation tools for investigating financial fraud detection systems' performance gaps are not in place. However, such tools are necessary in order to highlight missed areas to optimise further developments or investigational resources on. In this paper, an effective Visualisation and Clustering Tool based on Hierarchical Clustering is outlined. This tool aids the need of visualising the fraud detection systems decisions per observations across multiple benchmarks. The high-level overview guides the analyst to drill down in suspicious areas per possible mistake type of fraud detection systems. The potential afforded by combining diverse hierarchies and statistics of the fraud detection systems is presented in real-life scenarios. Additional capabilities such as heuristic clustering allow to dissect worse performing benchmark hierarchies in detail.

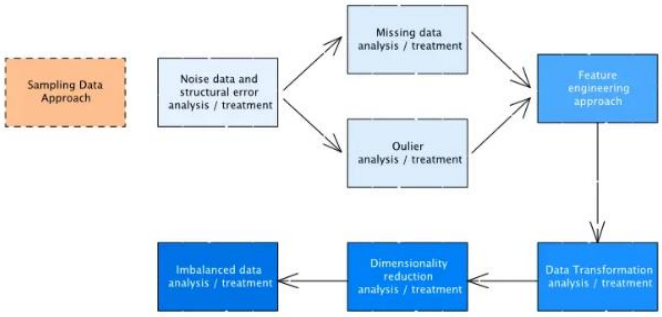

Fig 4: Data Preprocessing Techniques in Machine Learning

5.1. Data Cleaning

Data can be contaminated due to human errors, sensor errors, or unstable environments. The result of a data collection process can be misleading and contain a number of quality issues due to such contaminations. As analysts and scientists increase their reliance on this data, the consequence of dirty data is a more severe and harsher one in terms of the quality of the project. It can cause serious damage such as financial cost and damage trust in the systems with the dirty data. Therefore, a data cleaning stage should be considered in any data project. It prevents such possible issues from happening and guarantees accurate, relevant, and reliable information. Depending on the size and purpose of the data, a data cleaning solution should be considered and developed .

Data cleaning is to focus on correcting and removing errors in the data such as incompleteness, inconsistency, inaccuracy, irrelevance, and redundancy and transform the dirty data into better quality data. This stage has never been more important in the era of high performance and big data. Various approaches and algorithms have been developed to address an aspect of cleaning data including data validation, outlier detection and treatment, missing data imputation, and duplicate detection and treatment. Existing approaches vary from statistical tests to machine learning algorithms. Various approaches have also been developed further into products used by many companies or organizations. Though there have been plenty of contributions to improving the performance of cleaning, a lot of cleaning tasks still consume a substantial amount of time even with a thoroughly developed product. Some cleaning tasks such as outlier detection can cause long running time especially for big data sets. Despite the prohibitive running time, cleaning remains to be done by a human expert system and rule-based facilities to enable human interaction with a product. Such back and forth exchanges do not come free. In fact, human expertise is another common bottleneck during the analysis of data.

The result of a previous cleaning step may affect the next step so that the entire pipeline is sensitive to the decision made in a previous step. Many cases involve human assumption of which the data is more relevant to an analysis task. Such an assumption hard to justify would bias the entire analysis. It is so common that cleaned data sets have to be analyzed further and cleaned again before adopting a data set that is trustworthy. Depending on the application or target domain, a different definition of cleaning exists yet most existing approaches can clean (or evaluate) only one defined aspect of cleaning.

5.2. Feature Selection

Feature selection involves automatically generating a large number of potential predictive features from transactional data using the methods from transaction data mining. An important aspect of this setting is that huge caches of transaction data are available for each merchant for a long duration of time, thus creating a need for a scalable solution. Such problems have been addressed before in other application domains, with the focus being more on generating features exhaustively from the transactional data. Manual selection of features is prone to bias. Thus, the ideal setting that effectively mitigates

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

this issue is to automatically generate a large number of potential predictive features and then use a good selection method to pick the best subset. Ideally, the selection algorithm used in this process should also be able to scale well to large numbers of predictive features generated. A good feature selection method that can work on such a high dimensional dataset is Recursive Feature Elimination with Cross Validation (RFECV), which recursively removes the least significant features and thus can scale well to high dimensional datasets. RFECV can be effectively utilized through a distributed parallel approach since it is naturally a divide and conquer algorithm. Each slave worker can receive a subset of the original database and only work on this subset to remove some features, and this smaller resulting database can then be sent to the master worker. The master worker can then combine the results obtained from varying slaves and take appropriate action. Ordinarily, predictive features can be generated using a combination of both domain knowledge and statistical techniques such as hidden Markov models, Bayesian networks, or support vector machines.
However, since expert knowledge is hard to come by in this instance, an entirely automated feature generation technique was preferred. A good feature generation technique should be able to eliminate irrelevant features and should generate only a small subset of predictive features from the vast number of generated features. The Card Shark feature generation approach can be adapted for this purpose by employing restrictions on the number of separate transactions and also restricting the search space to use only aggregators supported on the prediction engines .

Equ : 3 Precision-Recall Optimization for Fraud Model

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}$$

- **Purpose:** Evaluates the balance between precision and recall for fraud models.
- **Variables:**
  β: Weighting factor (e.g., $\beta = 2$ to emphasize recall).
  Higher F_β indicates better model performance in detecting fraud.

5.3. Normalization Techniques

Data quality is an important issue for all learning algorithms. Therefore, data cleansing techniques such as removing redundant datasets are essential for increasing the efficiency of learning algorithms, the performance of a model is overwhelming for credit card dataset due to its huge features. The features of the credit card dataset are normalized by z-scoring. In general, distance based algorithms are sensitive to scale and distribution of values in a dataset. On the other hand, decision trees and tree ensembles are robust to scale and distribution of values in a dataset.

Big data including streaming datasets and hidden features should be processed carefully so that only the visible part can be extracted first. But in the phishing site dataset, the data distribution is relatively small (1500 records with 34 features). So, the features of phishing site are directly used without normalization, outliers are also directly used without any modification. On the whole dataset, a library called imbalanced-learn library is utilized to resample the dataset in order to tackle imbalance problems in the dataset. The library includes a collection of re-sampling techniques. Random under sampling technique is used to balance class distribution by randomly eliminating samples from the majority class. Some data instances of the majority class in the training set are randomly undersampled to balance the dataset. Synthetic Minority Over-sampling Technique is a widely used oversampling technique to create synthetic data instances of the minority class by interpolation.

VI. MODEL EVALUATION AND VALIDATION

Different tasks are specified for the final model evaluation. Each of these tasks ensures that the model is ready for production. Care is taken to evaluate it properly so that features and hyperparameters are not readjusted after each model training. Typically, after component feature selection and training, the initial model evaluation procedure is as follows. The data scientists first request a dataset in the same format as the day-ahead dataset, consisting of examples of the upcoming day where labels are not known. This dataset is used for predictions, but also to evaluate the deployment safety of the model. Immediately after retraining, the predictions on this dataset are calculated using a script that is also run in production. These predictions are then forwarded to the model evaluation notebook, which checks the model safety on the on-demand predictions (the upcoming day) against the snapshots used to train the model. The predictions are brought into production, and new incoming predictions are also checked to keep monitoring the model safety.

The expert errors that occur are noted. If a fraud case has been missed, the process is reviewed step by step to identify what caused this specific weakness. A result of this is that a matter that previously did not require attention is given more monitoring. All deployments and errors are registered in detail, along with their description and the data that is involved. This improves the understanding of what contributes to both positive and negative fraud detection. In turn, this insight improves production monitoring and enhances the model training.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

For instance, if it appears that fraud cases with a combination of product/tender id's are missed, they can be added to the special batch of cases that are stored, predicted, and monitored in detail.

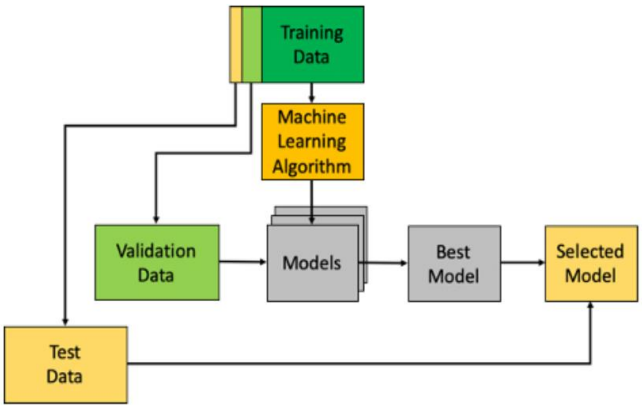

Fig 5: Model Validation Framework

Several validation methods are available to verify whether a model (or its specific components) satisfies requirements. These requirements are typically in natural language, which is vague, and it can often be quite difficult to verify the behaviour of a (black-box) ML model against the requirements. As a result, it is not uncommon for some requirements to remain unverified. The quality assurance framework requires that unit tests are formulated before the model is trained. Some require a bit of extra creativity to formulate, but other methods are already in use to comply with the requirements to some extent. Several types of checks take the trained model and input data from the day ahead and return true or false. These checks can be run on a standard model, and if it returns false, the model is excluded from use.

6.1. Performance Metrics

There are many aspects that should be monitored during the operation of the system, and new ones can be added as required. The measurement methods should not be susceptible to fraud and should have low operational costs all compared to the expected costs of the bad activity, where the starting parameters should include, for example, the relative risks and costs of ignoring such fraud. The key performance metrics is a measurement of the quality of the systems that checks transactions to provide a binary decision of a good transaction or a bad one. A unitary digit scale is suggested to be used in order to allow continuous fine-tune of the analysis. "Perfect" attendance of bad users as well as "Perfect" attendance of good users is impossible, and a minimum 95% and maximum 5% are allocated, respectively. The first measures the bad decisions and the second one the good ones which leads to collection of evidence and connection to authorities or termination of agents. Both measures of quality are needed during tests with their threshold of good value being set to >0.8. This value describes the relative risk versus cost of ignoring fraud behavior and will change in accordance to different situations, for example, if heavy delay penalties are imposed false positives will be additionally graded.

6.2. Cross-Validation Techniques

Validation in fraud detection is necessary. Sample data are almost always limited. The goal is primarily to know how well a model generalizes to a new dataset. The performance of a model fitted to one dataset may be assessed by training a model on some of that data, the training set, and estimating the performance on the left-out portion of that data, the test set. The trained model may be viewed as a sequence of approximations to the true model. A model may be selected from that sequence, i.e., a checkout of a value of the penalty parameter in some additive tree regularization framework. This means finding the model that, when fitted to the data, yields a sufficiently good performance on a new dataset, i.e., a model with a small fraud loss.

Availability of the fraction of fraudulent observations in the data is vital. Real fraud detection problems are typically characterized by a very small fraction of fraudulent observations. In such a situation, if the algorithms are trained on the original data, essentially all observations are of the non-fraudulent type. The model will then produce predictions that yield a large majority of zeros, and a high overall classification accuracy. However, this would not be a satisfactory fraud detection system. To alleviate the problem of imbalanced classes, synthetic observations, that is, new fraudulent observations, are produced, and the fractions of the two classes become more comparable.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

Tentative models are fitted to the data and the performance of those models assessed on a holdout set. Both metric and the data are ambiguous. They must be specified. Planning is needed, or a way to handle the ambiguity selected on the basis of model identification and estimation stability is needed. Integration across these stages is required. Data-dependent decisions made at training affect decision-making at test time. Models need to be fitted and cross-validated on training datasets that are custom-designed with respect to the test dataset, which may be unknown, and hence not accommodated. The uncertainty in the data is compounded by the uncertainty in how to validate models. The concern is fraud detection for the public sector. In particular, the detection of wrongful regulations of unwarranted reimbursements by principals.

VII. CASE STUDIES OF SUCCESSFUL IMPLEMENTATIONS

Many public sector organizations around the world are facing challenges in implementing analytics techniques to detect anomalies and frauds in public expenditure transactions. Stories of misappropriation of funds and unauthorized payments by public accountants and auditors related to enormous amounts of public money have been reported in many countries. Various supervisory and auditing actions are being employed to reduce illegitimate expenditures, generate alerts for the public administration in charge of them, and, if necessary, subject them to legal action. The findings of several domains, including economics, mathematics, statistics, accounting, fraud examination, and machine learning, may provide many and diverse practices for studying expenditure transactions of public organizations.

Big data technologies and analytics have been widely studied to detect aberrant behaviours in various domains. The capability to extract insights in real time from massive data streams, and to provide actionable intelligence with the goal of preventing mistakes, is crucial in public administration. Some frameworks based on distributed computing have been proposed in both public and private domains, but little attention has been paid to both refactoring and cluster deploying them. Moreover, the scientific community has developed a framework for the detection and visualization of anomalies in public expenditure transactions. This framework collects expenditure transactions of public agencies, classifies them, utilizes graph and metric data extraction techniques, and identifies significant anomalies based on interpretation rules.
Two case studies developed in the above-mentioned framework are presented below to demonstrate how it works to collect, analyze, and visualize the anomalies found in a real national dataset containing invoices of public accounting for public administration organizations. The Italian Ministry of Economy and Finances provides an application interface to access the Accounting DataBase. This database, called SIOPE, contains over 26 million invoices issued annually by about six thousand public agencies. The financial monitoring office provides a web service that allows stakeholders to extract invoices that have been or must be subject to monitoring. Some corporate public agencies are subject to ongoing monitoring by the Ministry of Economy and Finances, as well as to a prior payment prohibition. All other agencies are monitored after the payment takes place, according to a sampling procedure. In both scenarios, public administrators may be alerted to unexpected behaviours, and a black box strategy should be followed in order to not disclose the detection techniques.

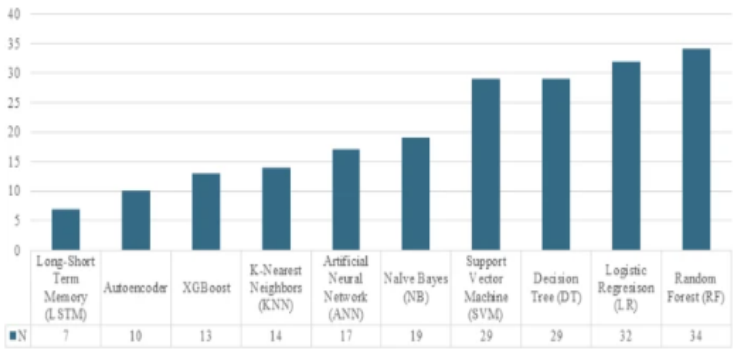

Fig 6: Financial fraud detection through the application of machine learning techniques

7.1. Case Study 1: Local Government

Without clear and accessible procedures, public sector credit cards can become a serious fraud detection issue waste taxpayer funds and degrade governmental and political credibility. For local governments, this case study examines how big data and machine learning (ML) might solve the problem of credit card expenditures that do not adhere to guidelines in a manner superior to extant methods. A simulation has been created to break down a local government's credit card pool by analyzing necessary data cleaning, and ETL processes, and thus facilitate fraud detection using random forests

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

with participation in care by local government capital cities. The scripts needed to accomplish this task have been created, which take 12-26 hours to run depending on the size of the city. Over the course of seven months, 1506 transactions that did not meet the analyses were flagged. Furthermore, fraud detection alone serves superfluous and irrelevant ends, and disclosure of excessive input information outside strict guidelines leads to greater risk for fraud. A suggestion is made to save the analysis results to accommodate the strictness of user specification clarification. Moreover, big data and ML can be used to rid the public sector of cranks that do not run similar analyses with supervised methods .

For any implementation, close supervision by financial oversight staff is suggested. Results from the previous seven months of local government credit card expenditures reveal a need for fraud detection systems based on data, because where fraud occurs within it economic feasibility decreases as more money is expended on it. From the latest analyses, 510 transactions have been analyzed, which have flagged discrepancies in credit card procedures as fraud regarding guidelines. Detecting additional inappropriate expenditures would require the additional devising of variables and careful thought, and thus user requirements and the extent of implementation must be weighed against its cost. For any implementation, the assembling of a statistical team consisting of staff with financial oversight who would devise a fraud system in cooperation with the financial department is critical. As suggested above, the implementation of big data systems to analyze financial transactions is needed.

The clean data used in the case study provide an idea of the potential size of such information, which is feasible for public sector agencies as they will be concerned with fraud detection. Credit card systems at the local government level frequently use the same software, and so the feasibility for public sector implementation is considered. One consideration is the potential forbidding cost of software to analyze the massive datasets to rid the public sector of cranks and flag questionable scores. The analysis also raises a number of critique points. There is a need for inner disclosure of stuff that must feasibly be hidden from the public, as it is transparent to investigators of card expenditures. Disclosures of input flags confounding factors make for greater chances of fraud detection.

7.2. Case Study 2: Federal Agencies

In the public sector, many federal agencies (FAs) are implementing or have implemented a new enterprise financial management system (EFMS) that integrates budgeting and accounting systems under one platform. The new systems have the opportunity to eliminate many long-standing discrepancies between budgets and actuals in agencies accounting for budgets/obligations/reserves on a cash basis because of their level of detail and their design for simultaneous execution of business transactions across multiple functions and systems.

The integration has the potential, therefore, to significantly impact the reporting of fraud-based outlays and the reporting of waste or extraneous obligations. This research describes the implications of adopting new EFMSs on financial fraud detection systems and then proposes a design for such a system powered by big data and machine learning techniques. To evaluate the design, a prototype system supporting budget vs actuals fraud detection is demonstrated using actual financial transactions from a federal agency's new EFMS. Findings confirm the novel opportunity of using big data and pre-trained generative ML systems to enhance fraud detection in federal agencies migrating to federal developed EFMSs. Event data may be leveraged to generate features for anomaly detection, while generative ML techniques may be applied to obtain inferences of frauds previously unknown to compliance leaders.

Many federal agencies (FAs) that disburse many hundreds of billions of dollars of appropriated funding each year are implementing or have implemented a new enterprise financial management system (EFMS) to replace legacy financial management systems. Financial management systems are the backbone behind how agencies account for and manage billions of dollars of appropriated funding each year. They capture the accounting events through which the budget is executed and address the budgeting/accounting/reporting of the results.

The new systems, the majority being cloud-based systems presently enhanced for federal use, have the opportunity to leverage the latest in artificial intelligence, machine learning and big data to significantly enhance financial fraud detection. Comprehensive reviews of conventional fraud detection systems identified the need for integration of financial and non-financial data in EFMSs in order for their fraud detection features to have enhanced effectiveness. Fraud detection in legacy systems was made more challenging by constraints that new systems do not have. Accordingly, there is a novel opportunity for the design and implementation of financial fraud detection systems that leverage big data and machine learning methods in a new EFMS.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

VIII. CHALLENGES IN IMPLEMENTING BIG DATA SOLUTIONS

Financial malfeasance continues to be a huge concern for the public sector as it hinders the effective and efficient provision of social services. As a consequence, several countries have implemented effective audit and control systems to ensure the proper management of public money. However, despite their adoption, fraud continues to occur at a significant scale, pointing to a limitation of current systems. For this reason, many current fraud detection systems are based on a compliance approach, which relies on process controls—a focus on what happened rather than on data. Nonetheless, with the advancement of technology in data generation and storage, this compliance approach is increasingly seen as inadequate to fully safeguard public sector financial systems.

In the spirit of "a picture is worth a thousand words," the public sector's need to use Big Data analytics to dig deeper into its data is becoming more evident. Despite its promise, adopting a Big Data approach for fraud detection is not an easy task, especially in sectors where data is sensitive. A review of the literature points to five main challenges when implementing Big Data solutions: (i) Threat to data privacy and security, (ii) Data critical for risk identification/assessment cannot be captured, (iii) Uncertainty of data, (iv) Inability to scale solutions across an organization, and (v) Data quality problems.

Data privacy and security are among the most critical challenges as the public-sector financial systems contain data that can be used for financial malfeasance itself, thus rendering the public sector attractive to malicious actors. Moreover, if safeguards around confidentiality and data access are not taken, Big Data has the potential to compromise data and defy regulations. Moreover, currently available secondary data is not able to capture qualitative aspects of public sector contracts, which are critical in identifying and assessing risks associated with these contracts. While several local government datasets have been made publicly available, these datasets are much more simplified compared to similar datasets that are kept private. Similar datasets are also inaccessible at the organizational level or at the national level.

8.1. Data Privacy Concerns

Fraud detection mechanisms greatly impact the stability of the financial sector, the protection of customer assets, and active engagement against money laundering. The research landscape of fraud detection in economic entities is understudied and unstructured, owing to the large variance in the applicable approaches. Investigating both large established financial institutions and young and agile fintech companies, three categories arise. Based on the presented results in the formal detailed analysis, numerous challenges are harnessed along with prospects for further future investigation. Supervised learning models on easy-to-interpret data yield the best precision in fraud detection. Digging deeper into data, unsupervised detection methods unveil conceptual pitfalls of an adversarial game. Furthermore, the privacy aspect is worth further exploration as the protection of customers is a revenue-generating priority.

Roughly executed qualitative clustering of fraud detection projects is valid both ways. Knowledge on fraud detection systems has been amassed and categorized thoroughly, offering structural clarity and an informative overview. The existing process pipelines can serve as templates for institutions developing new systems on detailed examples. Surprisingly, there are almost no substantial scientific publications on projects that date back a decade or longer, nor papers on the successes of existing systems. There is also an unaddressed trend in the dissemination of publicly available synthetic data sets for financial time-series problems, with a special emphasis on fraud.

New models and methodologies in machine learning have emerged to aid financial institutions in combating fraudulent activities, social engineering scams, and other illicit perceptions. There are well-established ecosystems of technology partners and risk management vendors servicing machine learning solutions to financial institutions. Applications of machine learning in fraud detection, transaction monitoring, Know Your Customer (KYC) control, and anti-money laundering efforts all exist. However, most of these tools are built on transaction logs within the ownership of a given financial institution. Many machines are built to catch ATMs, credit cards, or accounts behaving suspiciously, but what about actors unfamiliar to a given bank or payments processor?

8.2. Integration with Legacy Systems

Today's rapidly evolving financial landscape provides organizations with unparalleled opportunities. However, these opportunities are complemented by risks that must be effectively managed to protect organizations and their stakeholders. Financial risks faced by organizations relate to their financial systems for processing tax, auditing, and treasury. Financial fraud is on the rise within governments, resulting in loss of funds, maintenance of inaccurate data, damaged reputations, and lawsuits. Fraud detection systems can warn the organization about potential threats. Organizations can take appropriate action against fraudulent systems with blockchain platforms.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

Fraud detection systems heavily rely on vast datasets with various attributes and parameters to identify unusual, fraudulent claims related to public sector financial systems. Nowadays, the financial systems for budget processing, ledger posting, and bookkeeping for tax records have to ensure the analysis of the vast datasets to detect fraud using various mechanisms of AI techniques. The proposed architecture can be easily integrated with the existing financial systems without any modification; therefore, it can easily be adopted by organizations. The architecture has two main components: the BDS component at Cloud for finance data analysis using big data and ML and the BDSM component for processing finance unstructured data capturing crucial chartered accountant videos stored in archives to analyze and monitor them. The architecture offers a variety of services to its users depending on their needs.

Numerous analytical approaches can be used to analyze this extreme scale of data and to extract insights from them using methods of Big Data analytics. Finding statistical links between the different variables describing the input data might lead to useful insights. To this aim, it's necessary to look for regularities in how these input attributes relate to ground truth labels. How to match nD signals and data structures to extract analogous patterns using knowledge-driven or sophisticated machine learning and/or deep learning algorithms? From text data, how to automatically develop Taxonomy and Ontology models, which capture the semantics, for managing the understanding of the new vocabulary of domain words? .

8.3. Skill Gaps in Workforce

Some definitions that must be fitted: Monitoring Controllers: Often called accountants, controllers are responsible for investigating ecosystems and transactions in detail. They also define fraud prevention measures. Explorers: Responsible for business divisions, enterprise investigations and communications. Monitors and Explorers interact more than Explorers and Controllers. AI algorithms are developments such as measures against a previously unknown problem. They attract Explorers expertise. In the field of data science, there is a need for innovations that extend the analytical potential of existing consulting algorithms and mines to non-aligned data. Using it requires proper preparation of data scientists. Typical positions are Senior Data Scientist or Business Analyst. Mid-tier positions require strong advanced knowledge. Data analysts have a junior role in implementation and operations. They hold a BA-level IT and programming degree and do iterative and collaborative work.

These positions imply relevant degrees in engineering fields and languages that require the foundation. However, they are often placed at moderation levels. Problems can develop, such as producing misleading algorithms, too narrow views on analytics, or wasted talent in administrative jobs. The systemic data analytics knowledge would be narrowed down to classic statistics. Current education is not complemented by the knowledge expansion or option share. The workforce expansion should include business analytics and deeper big data. Cross-sectional understanding is needed between professional IT strengths, business drivers, and governance knowledge. Teaching programming, statistics, and formal analytical solutions would deepen the education's breadth. Existing data science programs often do not include both hard and soft skills. Though tools are often adopted, some targets are black-boxed.

Therefore, many professionals do not deeply know the potential of given packages or data mining algorithms. Increasing knowledge will develop adequate limits, reducing abuses and frustrations. Often, only statistical pre compatibility algorithms are used, neglecting the harsh service needed for complex tasks. Understanding classical statistics could be part of knowing commercial solutions. Trusting all data to packages could be naïve. Educating financially relevant software shaping externalities can increase knowledge. For example, their potentials, positive and negative, could be included in courses. Using it with a proper understanding could enhance users' benefit for the environment.

As introductory knowledge is necessary, a separate focus topic must be prepared for needed hard GM. Knowing IT easily and deeply is necessary for organized analytics. A deeper focus on technical development tools opposing business-focused education is scarce. The role of decision trees should be developed and abundant. Cross-sectional knowledge is desirable and enriching.

IX. CONCLUSION

Despite the rapid and revolutionary rise of AI and technological advancements throughout the years, fraud detection systems are still flagged as needing improvement and quick fixing. Several factors led to the difficulty in tackling this flaw in the financial IT management of public institutions, the most prominent being a lack of human resources in the finance sector. A human financial analyst can inspect an estimated 15,000 functions out of 40,000,000 daily transaction functions, given they work an 8-hour day, translating to an impossible amount of money laundering activities that can slip through the cracks.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 12, December 2020

DOI 10.17148/IJARCCE.2020.91221

At this scale, the traditional SQL-style counter fraud and money laundering check, though powerful, is incapable of capturing quickly developing new trends in fraud, with the potential to continuously evolve and propose a cascading army of human-like financial fraud prevention. Fortunately, with Big Data quickly coming together with more sophisticated machine learning and predictive analysis algorithms proposed by tech firms, the long desired raising of cost efficiency in the public sector financial IT systems could just be a step away.

While AI charmers provide the possibility for extreme cost-effective money laundering checks through deeper analysis and a greater volume of examined functions in lesser time, there are still key issues that need careful consideration and attention. General AI discussion proselytizes 21st-century digital slavery and global surveillance, with a growing tendency for authoritarianism surrounding the operation regulation of these technology companies. While the continuous black box issue resulting from rapid evolutionary processes hinders consumer responsiveness surrounding defenses against any issues concerning the application of sophisticated ML, these models are still being heralded as the new trend in technology and money-savvy checkers. Overall, the financial terms and analysis presented in clear and publicly accessible language come of immense assistance in understanding the landscape of money laundering checks and cyber fraud in public institutions.

Future Trends

When excessive query workload arrives, applications meet overload leading to performance degradation and possible service violation. Therefore it is important to assure the Quality of Service (QoS) of applications: response time stability (e.g. 95% hit rate not above 1s) and quality assured delivery (e.g. all queries should be delivered with success and within 60 s).

REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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.

[5] Ghahramani, M., Qiao, Y., Zhou, M., O'Hagan, A., & Sweeney, J. (2020). AI-based modeling and data-driven evaluation for smart manufacturing processes. IEEE/CAA Journal of Automatica Sinica, 7(4), 1026–1037. https://doi.org/10.1109/JAS.2020.1003114