Educational Administration: Theory and Practice
2024,30(10) 396- 407
ISSN:2148-2403
https://kuey.net/

Research Article

Leveraging Big Data And AI/ML For Fraud Detection In Retail Transactions

Vishwanadham Mandala[1*]

[1*]Data Engineering Lead USA, vishwanadh.mandala@gmail.com

Citation: Vishwanadham Mandala, (2024), Leveraging Big Data And AI/ML For Fraud Detection In Retail Transactions, *Educational Administration: Theory and Practice, 30(10), 396- 407*
Doi: 10.53555/kuey.v30i10.8103

ARTICLE INFO

ABSTRACT

Fraud is an intentional misrepresentation or deception that causes another person to act to his or her detriment. Fraud detection is a critical problem faced by many organizations, particularly in industries such as e-commerce, telecommunications, banking, and insurance. New developments in smart devices, new retail transactions, and the massive growth of these devices have introduced many new opportunities for fraudsters as well as challenges for organizations trying to combat fraud. Retail transactions containing personally identifiable information are now more prone to privacy concerns, data breaches, and fraud than ever before. Legacy fraud detection techniques of manually crafting rules and employing rules without real scientific analysis are ineffective against new retail transactions. Big data technologies provide industry-leading speed, performance, and scale insight. AI/ML models are showing superior performance for fraud detection tasks due to new developments in computer science and the availability of data streams. Retail transactions in the form of large-scale datasets generated in real-time can be tackled using big data technologies and in-database AI/ML. A novel and effective AI/ML-based fraud detection approach is proposed that employs big data and AI/ML concepts to safeguard organizations from fraudsters. Leveraging the distributed in-database capabilities of big data technology, fault-tolerant massive datasets can be processed in parallel and analyzed in seconds using AI/ML algorithms. The ingestion of massive datasets within the architecture leads to AI/ML models being employed in the big data technology environment, significantly enhancing speed, performance, and reliability. The approach consists of four models: 1. Data Sources - a detailed overview of data sources is provided, including simulated datasets with retail transactions. 2. Data Processing and AI/ML Architecture - a description of preprocessing tables and features for deployment in big data technologies and design of AI/ML architecture for modeling and deployment, where AI/ML models are embedded in SQL, fed by processed tables in data lakes, and scoring tables are created with predictions, is provided. 3. Results - the performance of the approach with four AI/ML models employed on simulated datasets is evaluated. 4. Conclusions and Future Work - a summary of the approach is provided together with the performance of the designed models, and several future work ideas are introduced. The approach is novel in its design and implementation using new technologies and datasets while focusing on retail transactions, and it is valid since robust performance is shown with diverse AI/ML models.

Keywords: Big Data Analytics, Fraud Detection, Artificial Intelligence (AI),Machine Learning (ML),Retail Transactions, Anomaly Detection, Predictive Analytics, Data Mining, Risk Management, Transaction Monitoring, Real-time Analysis, Fraud Prevention, Behavioral Analysis, Pattern Recognition, Data Integration, Synthetic Fraud Detection, Customer Insights, Algorithm Development, Automated Alerts, Security Analytics, Fraudulent Activity Detection, Deep Learning, Data Enrichment, Rule-based Systems, Neural

Networks, Data Visualization, Threat Intelligence, Scalable Solutions, Cross-channel Analysis, Risk Assessment.

1. Introduction

Financial fraud is one of the top concerns of every retail organization. Fraud is a wrongful or criminal deception intended to result in financial or personal gain. Fraud can occur in different forms, such as altering information in database records, scamming customers or the organization itself, and hacking the systems for money theft. Detecting and preventing fraud in financial transactions is a challenging task that involves massive amounts of financial data. Furthermore, fraudulent activities evolve over time, which impairs the performance of traditional detection techniques. Additionally, the online retail industry is a billion-dollar industry, which has also attracted criminal organization activities in the form of fraud and scams. Therefore, to maintain the integrity of the financial systems and enhance public trust in online transactions, various fraud detection techniques need to be established.

Fraud can occur in various forms and types, affecting individuals as well as organizations. It can lead to the loss of financial assets and sensitive personal information. Fraud detection involves examining data to discover and determine events that deviate from the expected behavioral pattern. Financial transactions can be defined as an exchange of monetary value either in cash form or in value form. With the rapid growth of the internet and electronic transactions, fraudulent activities are growing by leaps and bounds. Fraud in financial networks appears in various forms, such as loan fraud, insurance fraud, check fraud, retail transaction fraud, and money laundering. Fraud detection in financial transactions or networks is an intensively researched topic in academia and industry.

Among all the forms of fraud, fraud in retail transactions is the most common type occurring in the worldwide financial network. Fraud in retail transactions occurs when a retail customer purchases goods with a credit card and commits fraud. This type of fraud occurs by using a cloned genuine card, stolen card details with no physical possession of the card, or the use of a false identity card. With the rapid growth of the online retail market, fraud in retail transactions has also increased. Losses resulting from credit card and debit card fraud transactions are expected to reach significant amounts in the near future. Additionally, in the Asia Pacific region, retail fraud losses were reported to be substantial. Hence, there is an utmost need for detecting fraud in retail transactions and understanding the evolving pattern of fraud over time.

Fig 1 : The Role of AI and ML in Detecting Retail Fraud

1.1. Background of Fraud in Retail Transactions

Fraud in retail transactions has become an increasingly prevalent and sophisticated problem, posing significant financial and reputational risks to both retailers and consumers. Fraudulent transactions can occur across various retail settings, including brick-and-mortar stores and e-commerce platforms. The rapid digitization and proliferation of online shopping have further exacerbated the challenges in preventing and detecting fraudulent activities. Retail fraud encompasses a wide range of deceptive practices, with some of the most common types including payment fraud, account takeover, return fraud, coupon fraud, loyalty fraud, gift card fraud, and merchant fraud. Payment fraud is the most prevalent form of retail fraud, comprising card-not-present and card-present fraud. Account takeover fraud typically involves an unauthorized individual gaining access to a customer's account by exploiting weak passwords or identity theft. Return fraud occurs when an individual returns stolen or counterfeit goods for store credit or cash. Coupon fraud involves the misuse of discounts and promotions, while loyalty fraud exploits loyalty programs to obtain rewards without qualifying purchases. Gift card fraud encompasses any unauthorized obtaining and redeeming of gift cards, and merchant fraud involves retailers colluding with customers to complete fraudulent transactions. The implications of retail

fraud are far-reaching, affecting various stakeholders within the retail ecosystem. Retailers incur direct costs, such as chargebacks, penalties, and legal fees, as well as indirect costs, including lost sales, diminished brand reputation, decreased customer trust, a heightened need for compliance and auditing, and increased operating costs due to heightened security measures. Payment service providers face reputational damage, legal challenges, and regulatory scrutiny, while consumers may experience financial loss, emotional distress, and compromised privacy due to exposure to personal information.

1.2. Importance of Fraud Detection in Retail Transactions

As fraudsters devise new and more complicated methods, fraud detection in retail transactions is becoming increasingly important for merchants. Retailers are required to identify and minimize fraudulent activity while maintaining a smooth transaction process, protecting customer privacy, and preserving brand reputation. The process of identifying fraudulent credit and debit card transactions and determining an activity's legitimacy in light of an established set of rules or an individualized model is referred to as fraud detection. Identification of fraud takes place based on data from related transactions and is done either in batch mode or in real-time. Data collected about customer transactions or other data relevant to these transaction activities may not express fraud but rather genuine transactions; yet such data may often provide insight useful in fraud recognition and fall into a spectrum defined as a suspicious score. Fraud detection failures may produce a real loss or may produce accusations and denial of real business cases. Hence, card schemes and government authorities impose on the merchants liability policies that assess liability penalties to the merchants according to the number of frauds that they allow. Fraud detection, potentially even real-time detection, is important in preventing such losses.Historically, retail fraud detection systems have incorporated a fair amount of rule-based logic. Typically, at initial deployment, basic rules are applied concerning suspicious aspects of buying transactions. However, as the retail environment is dynamic, fraud patterns often change, and these rules might become outdated and less effective. As business case studies have shown, fraud detection rules require continuous refining and self-tuning in order to remain effective. Gradually, acknowledgment of this opportunity became widespread, and more merchants were actively seeking these types of solutions. A fraud detection system that has access to a dynamically updated repository of all transactions and, over a settled period of time, learns and adapts models that characterize normal, fraud-free daily activities. Initially, there is no knowledge about the specific fraudulent behavior, and the detection relies on preliminary models of genuine activities which are continuously adjusted. This paradigm fits well with the unknown analysis requirements and operational constraints of many retail environments. Bringing fraud cases to the knowledge of merchants comes in different levels of detail, from a simple suspicious score of a transaction to a full account or user profile with associated transaction history. As business integrity is an essential base for conducting healthy retail activity, it lies in the interest of numerous non-competitive merchants and acts as an enabler.

Equ 1: Logistic regression

1. Logistic regression with ±1 labels. Logistic regression (with ±1 labels) maximizes the likelihood

$$L(\beta_0, \beta) = \prod_{i:Y_i=1} p(X_i) \prod_{i:Y_i=-1} (1 - p(X_i)),$$

$$p(x) \triangleq \frac{1}{1 + e^{-(\beta_0 + \beta^\top x)}} = \frac{e^{\beta_0 + \beta^\top x}}{1 + e^{\beta_0 + \beta^\top x}}.$$

Show that this is equivalent to minimizing the cost function

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \log(1 + \exp(-Y_i(\beta_0 + \beta^\top X_i))).$$

Hint: Maximizing the likelihood is equivalent to minimizing the negative log-likelihood.

2. Big Data in Fraud Detection

Fraud detection has emerged as an important business task across many industries, driven by increased automation and online interactions. Banking, securities, and insurance were among the first industries to benefit from statistical fraud detection, followed by telecommunications, healthcare, and retail. E-commerce and social networking companies are relatively new but large players.As fraud and abuse detection is a diverse area of machine learning, similar tasks are related to anomaly or outlier detection. The work focuses on fraud detection in the context of retail transactions, using a specific example where cards are used to purchase petrol at gas stations. In this case, card use at gas stations is relatively rare, which is key for separating legitimate use from fraud. Forecasting legitimate transactions, using signal processing techniques, leads to a sample of suspected fraud cases, with the exception of some false negatives.Challenges arise, such as definition ambiguity: there are many types of fraud and fraudsters; model ambiguity: there are many possible models of legitimate use; and data ambiguity: some observations and a priori knowledge uncertainty. The approach taken is that of a combination of models, learned from data, with one aim being to reconcile soft detections with hard claims or statements.Advanced statistical modeling approaches using neural networks, radial basis functions, and their generalizations have been developed. Switching or regime models and mixture models have been implemented for soft decision modeling, detection, reconciliation, and claim modeling. These approaches exploit past observations for rationale and dynamism, modeling dependencies in time and data.Well-known and straightforward data processing tasks have been implemented, such as variable selection by relevance and redundancy elimination, data normalization, and selection of representative sample fractions for modeling.

Training samples have been tailored for various modeling approaches to ensure the maximum effectiveness of model learning. Models have been evaluated from different points of view and combinations of models have been run for various combinations.

Fig 2 : Big data Fraud Detection

2.1. Definition and Characteristics of Big Data

Big data generally refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. But just having a lot of data doesn't make it 'big'. In general, a collection of data is termed big data when it gets difficult to store, manage, process, or visualize it with traditional tools. Big data can be characterized by volume, velocity, variety, veracity, and value; also known as the 5Vs of big data. Volume refers to the amount of data. In big data analysis, data is continuously generated at a huge volume that cannot be handled or processed with traditional tools. It is estimated that a significant amount of data is produced every day, and the amount of data is growing exponentially. An analysis of social network data, for instance, may involve a billion nodes and trillions of edges. Velocity refers to the speed of data generation and processing. In big data analysis, data is continuously generated at a very high speed, much faster than what traditional tools can handle. For instance, a cellphone tower generates a large number of messages per second. Variety refers to the types of data. Big data can be structured, unstructured, or semi-structured, and traditional data management tools work only with structured data. Structured data refers to data that have a pre-defined data model, such as a database schema or fixed field lengths. Unstructured data does not have a fixed data model and is generally text or video data. Semi-structured data is a hybrid form of data and lies between structured and unstructured data. Veracity refers to the credibility of data. Data can be consistently inaccurate or incomplete, misleading or deceptive, and inconsistent or contradictory. There may be hidden patterns that are difficult and costly to detect. Such dubious data is called deceptive big data. It can lead to high-profile cases of wrongful arrest, wrongful convictions, prudent terminations, and bad credit scores if bias and prejudices of stakeholders are included in a predictive model. The value represents the usefulness of data. The real value of data is realized only after data processing, analysis, and visualization. It may yield important, useful, and actionable insights otherwise hidden that will dictate data-driven decisions such as fraud detection, customer segmentation, and stock market prediction.

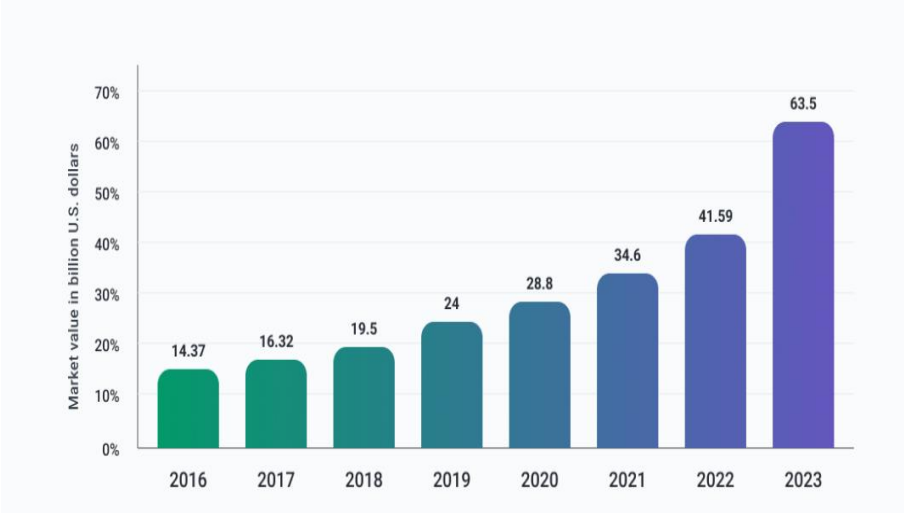


Fig : Graph Technology for Fraud Prevention

2.2. Challenges and Opportunities of Using Big Data in Fraud Detection

Despite the potential benefits associated with expanding the amount of available data, the significantly larger volume of data required for analysis raises processing and storage concerns. Five new, critical challenges are presented by the use of big data in detecting fraud. First, the integration of new data sources is a challenge because integrating multiple datasets drastically affects the computational time and the volume of the data, hindering the storage mechanisms. Second, the variety of big data types makes it challenging to easily analyze unstructured data. These types complicate the analysis process by requiring extra preprocessing steps before good quality analysis can be achieved. This difficulty is exacerbated when only one of the different datasets is used, which may produce analysis results with missing context. Third, the rapidly growing volume of data and the complexity of analysis methods create a technical knowledge gap; it is often complex for organizations to implement big data fraud detection systems due to the high acquisition and maintenance costs of analysis technologies and hiring of data scientists with extensive training and experience in big data solutions. Fourth, the constantly changing environment in which the analysis is performed means that the ability to remain up-to-date is pivotal. Specifically, many institutions have credit card fraud detection systems that prompt a review of transactions within seconds of completing a transaction. Such systems need to remain current to be effective; if a hacker plans to exploit a pattern of transactions overlooked, an outdated conceptualization would be vulnerable. Lastly, sensitive data is being used; the integration of possibly sensitive data sources introduces potential risks to the privacy of individuals. Thus, such integration should be carefully reviewed before being pursued.In addition, two challenges uniquely arising from the fraud context need to be considered. First, the cost of false alarms versus false dismissals in detecting fraud is highly asymmetric and severely context-dependent. Second, fraud is often highly imbalanced in its occurrence. Clearly stating and closely analyzing the specific fraud detection problem of interest is essential to avoid misleading and unusable results. Nonetheless, there has never been an exploration of the context under consideration, and this may lead to crucial issues in analyzing any fraud detection system. However, in addition to the challenges, big data nevertheless presents considerable opportunities in the context of fraud detection. The use of multiple unstructured data sources to gain additional context surrounding a potential fraud transaction may allow for a better understanding and more nuanced modeling of transactional behavior. Additionally, the ability to capture real-time data pathways through new sources encourages numerous opportunities for closer real-time monitoring of fraud behavior.

3. AI/ML Techniques for Fraud Detection

Artificial intelligence and machine learning have the capability to analyze a plethora of complex data, recognize patterns and anomalies to predict future events based on historical data, and are gaining broad attention in the detection of fraudulent activities. Automated systems using AI and machine learning are becoming a part of our everyday lives. AI and ML techniques, in conjunction with big data, are being used to cull pertinent information from financial transactions and customer history and erect models to detect fraudulent activities in real-time. Based on a set of variables, transactional information such as the location, amount, and type of transaction, models can be created to identify the normal behavior patterns of customers. With the help of these models, false transactions can be detected and flagged for further investigation. Fraud detection models continuously learn and adapt according to the category of the transaction or customer, intercepting targeted and advanced attacks.Support vector machines: A support vector machine is a supervised learning algorithm used to construct a model based on a dataset with features and classes to predict the class of unknown cases. The approach to construct the model is to define in a high-dimensional space a hyperplane that separates the classes in a way to maximize the distance to the nearest points. To enhance the estimated accuracy, instead of designing a single hyperplane, support vector machines design a set of hyperplanes looking for the one that best separates the classes in a voting scheme. This means that the more hyperplanes classify a case as belonging to a class, the more confident the prediction is. The methodology requires the optimization of parameters, which can be done with genetic algorithms. The support vector machine has shown promising results in a wide range of different applications.Decision trees: The decision tree is a supervised learning algorithm that can be used for classification or regression. The model is composed of nodes that represent the feature space and leaves that represent the predicted classes. Basically, this algorithm partitions the feature space recursively until predetermined stopping conditions are met or the partitions become pure, lying exclusively on a class. The recursive partitioning is conducted using the information gain concept that analyzes how much information and purity the new partition provides depending on the feature chosen and the value used to perform the split. Despite the concerns about overfitting, countable methods have been developed to enhance the complexity of the model, such as pruning and ensemble methods that combine several trees to create a more robust and accurate model.

Fig 3 : AI and ML in Fraud Detection

3.1. Overview of AI/ML in Fraud Detection

AI/ML refers to the use of artificial intelligence and machine learning algorithms to detect fraudulent transactions. In recent years, the use of AI/ML in fraud detection has gained popularity due to the increasing amount of data generated in retail transactions, the growing sophistication of fraudsters, and the need for real-time detection. Fraud detection using AI/ML typically involves two main steps: training a model on historical transactional data and applying the model to new transactions to predict the likelihood of fraud. The accuracy of the fraud detection system depends on the choice of algorithms, models, and features. Several different AI/ML techniques can be applied to fraud detection, including supervised, semi-supervised, and unsupervised techniques. In supervised techniques, a model is trained on a set of previously classified transactions (both legitimate and fraudulent) and used to classify future transactions. Supervised learning models typically include logistic regression, decision trees, support vector machines, random forests, and neural networks. Unsupervised techniques uncover patterns in transactional data without prior knowledge of a deposit fraud scheme. Unsupervised learning models, such as clustering and k-means, are typically employed to find abnormal or unusual transactions. The unsupervised approach is most suitable for retail transaction data fraud detection since only a small amount of data is labeled. Semi-supervised techniques can be employed when there is abundant legitimate transaction data available and only a small number of fraudulent transactions. Semi-supervised learning techniques build a model using supervised learning with a few labeled data samples and then improve the model using a larger set of unlabeled data samples. In the absence of pre-identified fraudulent transactions, moving average filters and anomaly detection techniques can be applied to identify potentially fraudulent activities.

3.2. Commonly Used AI/ML Algorithms in Fraud Detection

Several artificial intelligence (AI) and machine learning (ML) algorithms are commonly applied to fraud detection in retail transactions, with each offering unique advantages and limitations. This section presents the most widely used algorithms in order of technical complexity: logistic regression, decision trees, artificial neural networks, random forests, and clustering algorithms.

Logistic Regression Logistic regression is a statistical technique that examines the relationship between a dependent variable and one or more independent variables. By using logistic functions, this method calculates the likelihood of an event occurring, and it is frequently employed in binary classification scenarios, including fraud detection, where data is classified as either non-fraud or fraud. Logistic regression models can be easily implemented, described, illustrated, and tested for various data distributions. However, this method's main drawback is its reliance on the assumption that the relationship between the dependent variable and independent variables is linear.Decision Trees A decision tree is a flowchart-like structure in which internal nodes represent features, branches indicate decisions and leaf nodes correspond to outcomes. Decision tree algorithms recursively partition data into subsets based on the features' values until a stopping criterion is reached. The resulting tree can be intuitively represented and is simple to understand, making it suitable for cases with large datasets. However, decision trees tend to overfit training data and have high variance.

Artificial Neural Networks Artificial neural networks are composed of interconnected networks of simpler computational units called neurons, which can learn and perform functions by adjusting weights between connections. A neural network consists of an input layer, one or more hidden layers, and an output layer. Each layer has several neurons, and the output of one layer is passed as input to the next layer. Neural networks can learn complex patterns and relationships, making them well-suited to recognizing fraudulent transactions in retail transactions. Nonetheless, these networks require extensive computations, time, and iteration.Random Forests A random forest is a variant of decision trees that builds and merges thousands of trees based on bootstrap samples of training data and unique subsets of features at each tree split. Each tree supports a different decision for classification, and the final classification is determined by the majority vote. Random

forests reduce the overfitting issue of training data inherent in decision trees and have low variance. However, the main drawback is the reduced interpretability of the model.Clustering Algorithms Clustering is an unsupervised learning technique that groups unlabeled observations based on the distribution of underlying feature variables. K-means and hierarchical clustering are two clustering techniques frequently used as elementary statistical methods for fraud detection in transactions, often in combination with other classifiers. However, clustering algorithms can be computationally intensive for large datasets.

Equ 2: Anomaly Detection

$$MAD = median\left|x_i - \tilde{x}\right|$$ where $\tilde{x}$ is the median of the sample.

And the modified Z-score is defined as:

$$M_i = \frac{0.6745\,(x_i - \tilde{x})}{MAD}$$ where $E(MAD) = 0.675\,\sigma$

Iglewics and Hoaglin suggest that points with $M_i > 3.5$ be considered *outliers*.

4. Integration of Big Data and AI/ML in Retail Transactions

The explosive growth of technology is now enabling organizations to gather vast amounts of data about consumers' shopping habits, payment transactions, retail sales, social relationships, websites visited, and many other such activities. These transactions are being viewed through the lens of a vast and diverse data set commonly called big data. It is difficult to manage this data without a mechanism that helps store, retrieve, analyze, visualize, and infer predictively. AI/ML systems can be seen as a mechanism of choice for this purpose. The massive increase in unstructured and structured retail transactions these days has led to an increase in fraudulent transactions. In recent times, organizations have been observing considerable monetary losses due to poorly managed public safety and surveillance systems. Whether in online or offline shopping, credit card usage, payment mode, etc., there could be suspicious transactions of money through the theft or misuse of genuine user credentials. In addition, there are high chances of multiple transactions in quick time and with similar data usage. There is a necessity for efficient deployment of big data platforms for online real-time analysis and detection of these types of fraudulent and suspicious transactions. In recent years, many credit card transaction companies have been studying cardholder transactions and implementing various techniques to recognize patterns of valid transactions and detect fraudulent transactions. This, however, is not an easy task. Despite many precautions taken, it is impossible to stop or prevent fraudulent credit card transactions, as everyone has the right to accept the case of being cheated. Nevertheless, the detection of fraud is of utmost importance, and many different techniques have been proposed to detect fraudulent credit card transactions. Major challenges include false positives and false negatives. On the one hand, if too many false positives occur, the cardholder may be erroneously denied a transaction; on the other hand, if too many false negatives occur, the credit card company may lose considerable amounts of money.

4.1. Benefits and Limitations of Integration

As retail transactions generate massive amounts of data that can be leveraged for AI/ML fraud detection tools, it is necessary to integrate big data technologies with AI/ML frameworks. Integration of big data technologies and AI/ML frameworks in retail transactions provides a number of benefits. Integration of big data technologies with AI/ML frameworks reduces the time taken to train ML models and makes them suitable for real-time fraud detection. Technologies can be used for distributed data storage and processing and can be used for distributed training of ML models. Integration of fraud detection tools with various big data technologies also allows the processing of multi-dimensional datasets, including structured, semi-structured, and unstructured data from various sources, such as transaction logs, blacklists, and social media. Integration ensures that multi-dimensional datasets can be pre-processed, refined, and transformed into formats suitable for AI/ML fraud detection tools in a uniform manner. Can be used to extract, transform, and aggregate datasets from various sources for use in AI/ML modeling. Differentiating datasets based on the temporal or non-temporal nature of data can also be done uniformly. Datasets containing anomaly cases can be selected with a specific tolerance for AI/ML fraud detection tools. Such analytics and preventive actions for various incidents can be demonstrated using dashboards for decision-support systems. Several frameworks for integration of big data technologies with AI/ML or data analytics frameworks have emerged that offer tools for seamless integration. Provides HBase, Hive, and R integration; can be integrated with R libraries. Integrating commercial big data technologies with open-source AI/ML or analytics tools is costly, and there are limitations such as the need for third-party expertise. Tools for seamless integration are expensive. Integration between several big data technologies and AI/ML tools is still in the nascent stages. There are SQL-like languages for programming AI/ML algorithms on this basis. The main challenge is to build flexible, open-source big data technologies in parallel with AI/ML frameworks. The integration of big data technologies with AI/ML fraud detection tools is also possible in a corporate IT environment by developing some extensions and wrappers

around AI/ML scripts. Such systems for integration would contain components for transformations, data sampling, formatting, clustering, and proof of concept applications.

Fig 4 : Benefits of Financial Fraud Detection Using Machine Learning

4.2. Case Studies and Examples

With the growth of digital payments, the retail industry is becoming a prime target for financial crimes such as money laundering, identity theft, card-not-present fraud, and more. Retailers receive large volumes of transactions on a regular basis, making data-driven risk assessment challenging. Understanding transaction data is complex, involving multiple entities and factors, such as the point of sale and terminal ID, transaction value and currency, customer account, merchant, merchant authentication method, account holder IP address, and physical location of the terminal. Given the real-time nature of retail transactions, patterns indicating fraud are often affected by external factors, rendering previously trained methods obsolete.

Fortunately, the emergence of big data technologies capable of delivering speed, variety, and volume characteristics required for collecting, storing, processing, and serving data from different sources has transformed the situation. This has become possible with the growth of the internet and smartphones. Retailers are developing applications based on big data technologies to aggregate transaction data from multiple sources and applications to obtain a deeper understanding of their information and business processes. Data lakes created on big data technologies allow businesses to combine their data with public, partner, and third-party information. Analytical approaches utilizing various AI/ML techniques for real-time analysis of big data streams are being developed. With the advancement of the IoT, connected devices that can generate a variety and huge volume of information in real-time are becoming widely used in retail, such as payment terminals, card readers, parking meters, tolls, gas pumps, ATMs, and various wearables.Major financial companies have implemented systems to detect fraudulent transactions in real time using AI/ML models. These systems have delivered an ROI exceeding 1:10 within less than six months. In 2017, multiple machine learning models based on big data technologies for transaction fraud detection were developed as part of a solution for a payment service provider. The resulting solution was Benchmark-1 and earned a certification award from a research unit.The rapid expansion of digital payments has made the retail sector increasingly susceptible to financial crimes such as money laundering, identity theft, and card-not-present fraud. The sheer volume and complexity of transaction data—encompassing elements like point of sale, transaction value, customer accounts, and device information—pose significant challenges for traditional fraud detection methods, which can quickly become outdated. However, the advent of big data technologies has revolutionized this landscape. By leveraging data lakes and advanced AI/ML techniques, retailers can now aggregate and analyze vast amounts of transaction data in real time, integrating insights from public, partner, and third-party sources. This capability is further enhanced by the Internet of Things (IoT), with connected devices generating a wealth of real-time information. Financial companies have successfully implemented these technologies to detect fraudulent transactions with impressive results, achieving an ROI exceeding 1:10 within six months. Notably, in 2017, a payment service provider developed Benchmark-1, a machine learning-based fraud detection system, which earned certification for its innovative approach and effectiveness.The rapid rise of digital payments has significantly heightened the vulnerability of the retail sector to financial crimes such as money laundering, identity theft, and card-not-present fraud. The vast and intricate nature of transaction data—including elements like point of sale, transaction value, customer accounts, and device information—has made traditional fraud detection methods increasingly ineffective. However, the advent of big data technologies has transformed this landscape, enabling retailers to aggregate and analyze extensive volumes of data in real-time. By utilizing data lakes and sophisticated AI/ML models, retailers can integrate insights from diverse sources, including public and third-party data, to enhance fraud detection capabilities. The integration of IoT devices further amplifies this capability, providing a constant stream of real-time information. Financial companies have successfully harnessed these technologies, achieving remarkable results with an ROI exceeding 1:10 within six months. In 2017, a notable achievement in this field was the development of Benchmark-1, a machine learning-

based fraud detection system by a payment service provider, which received certification for its innovative and effective approach.

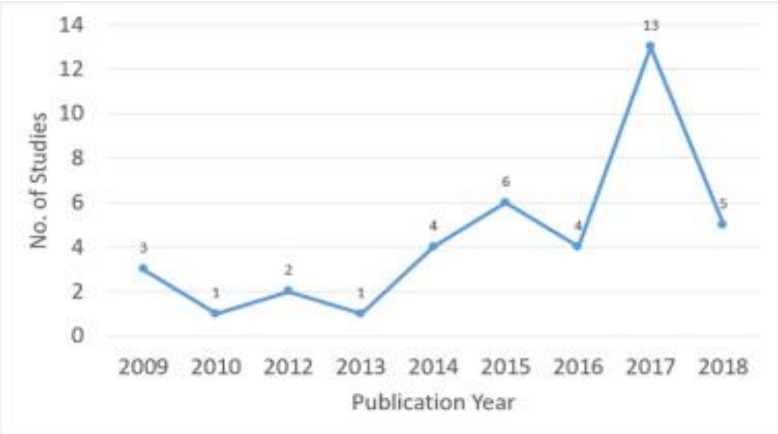


Fig : Fraud detection

5. Future Trends and Implications

In the coming years, organizations will accelerate their use of advanced technologies to detect and prevent fraud in retail transactions. Companies will bring new fraud detection use cases into production or enhance existing implementations with one or more of these capabilities. Organizations will become more skilled at evaluating their data environments, identifying opportunities for fraud detection advancements, and prioritizing feasible use cases. The use of advanced technologies such as AI, machine learning, IoT, and blockchain will become mainstream for fraud detection in retail transactions. With growing capabilities in fraud detection technologies, organizations will increasingly partner with technology providers to enhance their analytics and modeling capabilities. Vendors will face increasing pressure to deliver advanced product functionalities for monitoring and auditing their AI models. The need to ensure model interpretability and accuracy is paramount when deployed into mission-critical systems, especially for use cases such as fraud detection and behavioral analytics, where the potential harm caused by model errors will significantly impact the end customer. Blockchain will enable retailers to track the transaction history of products transparently to their customers, preventing significant losses caused by counterfeit products and opportunities for other fraud schemes. Creating new standards and regulations specifically for blockchain technology in retail transactions will be key for encouraging its use in the future. Technologies such as social media analysis, i.e., finding transactions with common users who have previously flagged behavior, will become more popular and focused on in the future, as they allow for finding impossible-to-detect fraud schemes otherwise. Furthermore, organizations will increasingly become aware of the opportunities for collaboration to combat fraud across boundaries, as fraud schemes become more sophisticated and international. These future trends and implications can be categorized into three main themes. The first theme involves characteristics of the emerging technologies that are anticipated to have a significant impact, the second theme addresses potential obstacles to fully realizing their positive anticipated impacts, and the final theme involves requirements for enablers to better secure their positive anticipated impacts.

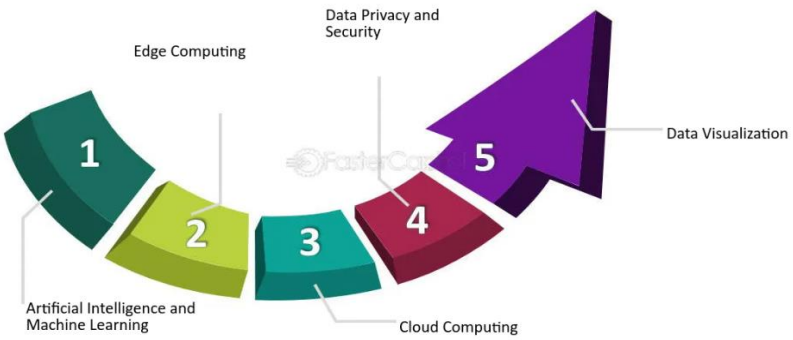


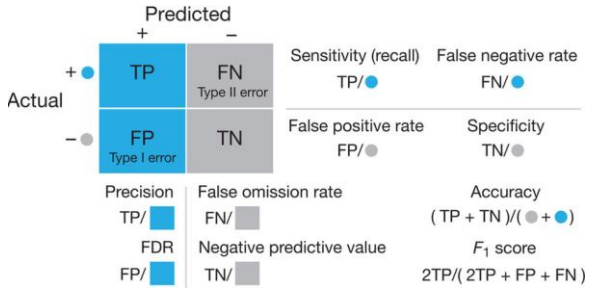
Fig 5: Future Trends In Data Analytics And Its Impact

5.1. Emerging Technologies in Fraud Detection

To detect fraudulent transactions in retail banks, big data analytics is hugely beneficial as large amounts of transaction datasets and user-generated datasets can be analyzed very effectively in a short span of time. Various organizations have started using big data analytics to detect fraudulent transactions. Several

algorithms are implemented over large datasets to detect fraudulent transactions in retail banks, telecommunications companies, e-commerce shopping websites, etc. It discusses various fraud detection techniques that are very effective in detecting fraud in their respective domains. Organizations have also started using handwriting recognition systems, as well as face and voice recognition systems to prevent identity fraud. Various automatic systems and tools have been developed in recent years that help to detect fraud very efficiently. Video surveillance is also one of the effective techniques to detect fraudulent activities, especially in retail shops. This method observes the illegal activities occurring in the shops and reports them to the system. The presence of large amounts of datasets offers interesting opportunities for automatic decision-making processes that allow the discovery of new fraud patterns. With the vast growth of the World Wide Web and various types of online services, fraud detection has become an important issue in the worldwide scenario. To counter fraud, there is a need for appropriate systems capable of detecting unauthorized intruder activities on large-scale systems with high accuracy while handling the enormous volume of new data being generated within very short periods of time.Machine learning algorithms, when combined with big data technologies, can be used to tackle the difficulties in fraud detection due to the vast volume of datasets. The diverse datasets and high dimensionality of datasets can be resolved using big data technologies, which offer an efficient way to store and analyze large amounts of datasets. The framework has its own data storage system and parallelism data computations. Fraud detection in retail transactions is a supervised classification problem that predicts whether the transaction provided as input is fraudulent or not based on the patterns captured by machine learning algorithms from the prior historically labeled datasets. To detect fraud in retail transactions efficiently, associated methodologies including data preprocessing, sampling, and post-processing methods also play a crucial role. The potential of big data analytics is exploited and applied for various fraud detection domains and banks across many countries. A significant amount of research has been carried out in fraud detection in retail transactions, credit cards, e-commerce shopping websites, etc., along with in-depth details of their own algorithmic procedures and datasets. However, there is a need for a comprehensive survey paper that focuses on big data and AI/ML technologies for fraud detection in retail transactions, and credits have not yet been brought into the limelight.

Equ 3: The confusion matrix shows the counts of true and false predictions



6. Conclusion

Fraud cannot be eradicated completely. However, it can be reduced to levels that are tolerable for all parties concerned. Robust fraud detection models require investments and resources. It is impossible to prevent fraud with 100% certainty. Cost-benefit analysis is necessary to find an equilibrium point where the cost spent to curtail fraud is less than the damage incurred due to fraud. Detection of Fraud is a non-trivial Industrial Big Data challenge but one that can have huge pay-offs once effectively solved. A comprehensive solution based on the use of novel approaches, tools, and technologies offered by Big Data is proposed for the detection of Fraud in Retail Transactions. A Big Data platform is built, and tools/applications are developed for Data Ingestion, Data Processing, Data Analysis, and Data Visualization. Various types of Data and Techniques are deemed necessary to build a comprehensive solution. The Big Data platform backed by the developed tools/applications allows the Implementing Company to ingest, process, and analyze Large Volumes of Internal Retail Transaction Data, complemented by External Data Volumes, Variances, and Types. A Modular Approach is recommended to realize all the proposed methods and tools/applications. Various Novel Data Sources External and Internal to Retail and E-Commerce Enterprises are Detailed within this Approach. Such data sources allow the detection of Temporal and Demographic Patterns in the retail sector. The combination of Big Data with new sources and forms of Data is, arguably, the Novelty and Competitive Advantage of the Proposed Solution. A prototype application is successfully developed, demonstrating the processing of Big Data using a Complex Framework. Statistical Analysis, Supervised Machine Learning Classification, and Probabilistic Graphical Modeling Techniques are proposed for Detection of Fraud in Retail Transactions. Such detection is realized via Novel Hybrid Models that Integrate More than One of the Proposed Techniques. Future Trends: Internally and Externally, the Business Environment is Changing and will Influence Online and Offline Industrial Fraud Detection. The Internet of Things and Wearable Technology Empower and Complete E-Transactions. A Cure for Noise and a More Complicated Web of Transactions. Proliferation of Internally Greater Diverse and More Ubiquitous Data Sources. The Emergence of New Payment Methods. A Proliferation of Drones is a Danger for

New Fraud Attacks or New Mine for Fraud. Characters a New Paradigm for Security and Fraud Protection or, Its Inflexibility, Create New Types of Fraud Capitalizing on Structured Information. External Factors May Induce Unforeseen Fraud Modifications. This Uncertainty Forces Constant Monitoring of Fraud Detectors and Attempts at Variable Speed Alterations of Internal Factors Varied and Vague. Constant Complex Calculations Outside the Oversight and Control of Decision-Makers. Automated Responses Execute a Predefined Set of Actions in Case of a Detected Fraud Type.

6.1. Future Trends

The future of big data, AI, and machine learning holds immense potential to transform the capabilities of fraud detection systems in retail transactions. There is a growing need for such systems to ensure trust and reliability in e-commerce. As the industry shifts towards digitization, conventional data, and model management techniques need to evolve to accommodate streaming data with increased velocity and volume. Near-real-time analytics systems are needed to store decision-making models and verify transactions as they take place. With the development of distributed data storage systems, graph processing, and big data analytical systems, the means to respond to this demand are here. Further research in this area can guarantee a better, safer, and more user-friendly shopping experience for both retailers and customers.

With the rapid development of machine learning techniques, fraud detection systems can further evolve by applying unsupervised learning or deep learning algorithms that can derive complex patterns from logs rather than relying on manually defined rules. Recently, convolutional neural networks for graph-structured data have emerged to model complex relationships between different entities, which can be trained on graph data in an unsupervised way. Because of the particularity of fraud detection systems, their application needs to be studied as well as the necessary model modifications to make them usable on data with very large cardinalities.The use of modern white-box interpretable models makes it possible to explain even complex predictions. In fraud detection, such means can be of utmost importance for investigators to understand why a particular decision was made either by the model or the analyst. This can help identify new strategies on both sides, fraudsters and the entity attempting to minimize the loss. However, it needs to be guaranteed that certain sensitive information does not leak to outside entities. The ability to make a model robust to certain transformations has been reached even for models as complex as neural networks, but it needs more analysis to establish what kind of transformation it can hold safely. For fraud detection systems with sound explanations, it is particularly important to know how robust they are to the adjustments and future changes in the monitored transactions.Because of the constant effort on the part of fraudsters to evade detection, there is a constant need for fraud detection systems to adapt to changing conditions. Therefore, the research literature on the use of active learning strategies is becoming more relevant to the field of fraud detection. The very novel technique that can autonomously extend its own knowledge without or with very little human effort is a universal agent, but its applicability to fraud detection needs to be investigated.

7. References

1. Avacharmal, R., Pamulaparthyvenkata, S., & Gudala, L. (2023). Unveiling the Pandora's Box: A Multifaceted Exploration of Ethical Considerations in Generative AI for Financial Services and Healthcare. Hong Kong Journal of AI and Medicine, 3(1), 84-99.
2. Vaka, D. K., & Azmeera, R. Transitioning to S/4HANA: Future Proofing of Cross Industry Business for Supply Chain Digital Excellence
3. Manukonda, K. R. R. (2023). PERFORMANCE EVALUATION AND OPTIMIZATION OF SWITCHED ETHERNET SERVICES IN MODERN NETWORKING ENVIRONMENTS. Journal of Technological Innovations, 4(2).
4. Mandala, V., & Kommisetty, P. D. N. K. (2022). Advancing Predictive Failure Analytics in Automotive Safety: AI-Driven Approaches for School Buses and Commercial Trucks.
5. Chintale, P. (2020). Designing a secure self-onboarding system for internet customers using Google cloud SaaS framework. IJAR, 6(5), 482-487.
6. Kommisetty, P. D. N. K., & Nishanth, A. (2024). AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI. In IARJSET (Vol. 9, Issue 10). Tejass Publishers. https://doi.org/10.17148/iarjset.2022.91020
7. Avacharmal, R., Sadhu, A. K. R., & Bojja, S. G. R. (2023). Forging Interdisciplinary Pathways: A Comprehensive Exploration of Cross-Disciplinary Approaches to Bolstering Artificial Intelligence Robustness and Reliability. Journal of AI-Assisted Scientific Discovery, 3(2), 364-370.
8. Vaka, D. K. (2024). Enhancing Supplier Relationships: Critical Factors in Procurement Supplier Selection. In Journal of Artificial Intelligence, Machine Learning and Data Science (Vol. 2, Issue 1, pp. 229–233). United Research Forum. https://doi.org/10.51219/jaimld/dilip-kumar-vaka/74
9. Manukonda, K. R. R. Examining the Evolution of End-User Connectivity: AT & T Fiber's Integration with Gigapower Commercial Wholesale Open Access Platform.
10. Kommisetty, P. D. N. K., & Abhireddy, N. (2024). Cloud Migration Strategies: Ensuring Seamless Integration and Scalability in Dynamic Business Environments. In International Journal of Engineering

and Computer Science (Vol. 13, Issue 04, pp. 26146–26156). Valley International. https://doi.org/10.18535/ijecs/v13i04.4812
11. Mandala, V., & Mandala, M. S. (2022). ANATOMY OF BIG DATA LAKE HOUSES. NeuroQuantology, 20(9), 6413.
12. Chintale, P. SCALABLE AND COST-EFFECTIVE SELF-ONBOARDING SOLUTIONS FOR HOME INTERNET USERS UTILIZING GOOGLE CLOUD'S SAAS FRAMEWORK.
13. Avacharmal, R., Gudala, L., & Venkataramanan, S. (2023). Navigating The Labyrinth: A Comprehensive Review Of Emerging Artificial Intelligence Technologies, Ethical Considerations, And Global Governance Models In The Pursuit Of Trustworthy AI. Australian Journal of Machine Learning Research & Applications, 3(2), 331-347.
14. Vaka, D. K. (2024). Procurement 4.0: Leveraging Technology for Transformative Processes. Journal of Scientific and Engineering Research, 11(3), 278-282.
15. Kodanda Rami Reddy Manukonda. (2023). Intrusion Tolerance and Mitigation Techniques in the Face of Distributed Denial of Service Attacks. Journal of Scientific and Engineering Research. https://doi.org/10.5281/ZENODO.11220921
16. Kommisetty, P. D. N. K., & dileep, V. (2024). Robust Cybersecurity Measures: Strategies for Safeguarding Organizational Assets and Sensitive Information. In IJARCCE (Vol. 13, Issue 8). Tejass Publishers. https://doi.org/10.17148/ijarcce.2024.13832
17. Perumal, A. P., & Chintale, P. Improving operational efficiency and productivity through the fusion of DevOps and SRE practices in multi-cloud operations.
18. Avacharmal, R. (2022). ADVANCES IN UNSUPERVISED LEARNING TECHNIQUES FOR ANOMALY DETECTION AND FRAUD IDENTIFICATION IN FINANCIAL TRANSACTIONS. NeuroQuantology, 20(5), 5570.
19. Vaka, D. K. (2024). From Complexity to Simplicity: AI's Route Optimization in Supply Chain Management. In Journal of Artificial Intelligence, Machine Learning and Data Science (Vol. 2, Issue 1, pp. 386–389). United Research Forum. https://doi.org/10.51219/jaimld/dilip-kumar-vaka/100
20. Kommisetty, P. D. N. K., vijay, A., & bhasker rao, M. (2024). From Big Data to Actionable Insights: The Role of AI in Data Interpretation. In IARJSET (Vol. 11, Issue 8). Tejass Publishers. https://doi.org/10.17148/iarjset.2024.11831
21. Reddy Manukonda, K. R. (2023). Investigating the Role of Exploratory Testing in Agile Software Development: A Case Study Analysis. In Journal of Artificial Intelligence & Cloud Computing (Vol. 2, Issue 4, pp. 1–5). Scientific Research and Community Ltd. https://doi.org/10.47363/jaicc/2023(2)295
22. Kommisetty, P. D. N. K., & Nishanth, A. (2024). AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI. In IARJSET (Vol. 9, Issue 10). Tejass Publishers. https://doi.org/10.17148/iarjset.2022.91020
23. Perumal, A. P., Deshmukh, H., Chintale, P., Desaboyina, G., & Najana, M. Implementing zero trust architecture in financial services cloud environments in Microsoft azure security framework.
24. Avacharmal, R., & Pamulaparthyvenkata, S. (2022). Enhancing Algorithmic Efficacy: A Comprehensive Exploration of Machine Learning Model Lifecycle Management from Inception to Operationalization. Distributed Learning and Broad Applications in Scientific Research, 8, 29-45.
25. Vaka, Dilip Kumar. "Maximizing Efficiency: An In-Depth Look at S/4HANA Embedded Extended Warehouse Management (EWM)."
26. Kommisetty, P. D. N. K. (2022). Leading the Future: Big Data Solutions, Cloud Migration, and AI-Driven Decision-Making in Modern Enterprises. Educational Administration: Theory and Practice, 28(03), 352-364.
27. Manukonda, K. R. R. (2023). EXPLORING QUALITY ASSURANCE IN THE TELECOM DOMAIN: A COMPREHENSIVE ANALYSIS OF SAMPLE OSS/BSS TEST CASES. In Journal of Artificial Intelligence, Machine Learning and Data Science (Vol. 1, Issue 3, pp. 325–328). United Research Forum. https://doi.org/10.51219/jaimld/kodanda-rami-reddy-manukonda/98
28. Perumal, A. P., Deshmukh, H., Chintale, P., Molleti, R., Najana, M., & Desaboyina, G. Leveraging machine learning in the analytics of cyber security threat intelligence in Microsoft azure.
29. Muthu, J., & Vaka, D. K. (2024). Recent Trends In Supply Chain Management Using Artificial Intelligence And Machine Learning In Manufacturing. In Educational Administration Theory and Practices. Green Publication. https://doi.org/10.53555/kuey.v30i6.6499