Journal of Technological Innovations

Est. 2020

Fraud Detection in Banking Applications: Machine Learning Approach

Khirod Chandra Panda
Asurion Insurance, VA, USA
0009-0008-4992-3873
Email: khirodpanda4bank@gamil.com

Abstract:

Fintech, a burgeoning sector in the global market, increasingly relies on IT to automate financial services, enabling around-the-clock, easily accessible transactions for a global customer base. This growing trend highlights the pressing need for security, as these information-rich transactions are prime targets for fraud, causing significant harm to both consumers and service providers. In addressing these security concerns, Machine Learning (ML) methods are employed to detect anomalies and predict fraudulent activities in Fintech platforms. Our contribution to this field includes a thorough evaluation of various anomaly detection techniques, analyzed across multiple real and synthetic datasets related to fraud. Our findings verify the variable success rates of ML in fraud detection, which led to a deeper investigation of the impact of specific features on the performance of these methods. Additionally, the paper analyzes the broader implications of these findings for the future security of Fintech services, acknowledging the high stakes of vulnerabilities that can lead to considerable financial and reputational losses. By proposing a machine learning approach, we aim to advance fraud detection and develop strategies to mitigate and recover from such fraudulent incidents. The core of our research involved examining a plethora of intelligent algorithms on a publicly available dataset, which was adjusted to reduce class imbalances, thereby enhancing the precision of our proposed algorithm. The results of this comprehensive analysis may accelerate verification processes and reduce the incidence and impact of fraud in the banking sector

Keywords: Fraud detection; machine learning; anomaly detection; Fintech; cybercrime,

1. Introduction

As the tempo of modern life accelerates, so does the demand for continuous, globally accessible services. The financial sector has seen a paradigm shift with the advent of Financial Technology, or Fintech, which blends IT innovation with financial expertise to offer enhanced services. Despite ongoing debates around its definition, Fintech has undeniably become synonymous with the progressive integration of technology and finance. The emergence of Fintech is evidenced by a surge in investments, attesting to its growing appeal among consumers and service providers alike, and positioning it as a potential successor to traditional financial institutions.

Fintech's allure partly lies in its diverse payment systems, including credit transactions and digital currencies anchored in blockchain technology. These systems enable seamless financial exchanges, but their reliance on IT infrastructure makes them vulnerable to exploitation through fraudulent activities that can have devastating repercussions. Acknowledging this risk, the paper highlights the adoption of machine learning (ML) for anomaly detection, identifying potential fraud within financial networks.

The paper examines the use of ML in generating predictive models from datasets, outlining the inherent challenges—most notably, the balance between false positives and service quality. The efficacy of various ML techniques is explored through real and synthetic datasets, showcasing their respective capabilities and

limitations in fraud detection within the Fintech domain.

Looking ahead, the implications of these findings on the future security of Fintech applications are significant. As the banking sector evolves, incorporating groundbreaking technologies like blockchain, AI, and big data, it confronts the dual challenge of innovation and security. This transformation is driving banks to adopt novel tech solutions, resulting in a shift from traditional banking operations to more customer-centric models.

The paper proceeds to delve into the sweeping changes across banking functions, emphasizing how technology has revolutionized customer service, yet has also heightened the sector's vulnerability to financial crises. Fintech's innovative stride has been instrumental in bridging the gap between customer expectations and the realities of banking services.

AI's emergence as a cornerstone in the banking ecosystem marks a move towards more intelligent, user-friendly

services. Advanced technologies, including AI-powered chatbots and biometric identification, are reshaping the way banking services are delivered. Yet, this evolution comes at a cost, as traditional banking jobs are at risk, underscoring the need for a workforce equipped with a new set of skills tailored to the demands of Fintech.

Fraudulent transactions remain one of the most pressing issues for banks, with substantial financial and reputational consequences. It is a growing challenge that necessitates the development of sophisticated fraud detection models suitable for banking institutions of any size.

This research endeavors to create a smart, machine-learning-based system capable of predicting and adapting to fraudulent behavior in banking transactions. The paper is methodically structured into sections covering a literature review of existing research, the impact of technology on banking in India, the role of AI in risk management, and fraud analysis through machine learning, culminating in a conclusion that ties together the insights gathered.

2. Literature Review

Statistical techniques have been leveraged for the identification of fraudulent activities by examining the statistical distribution of data for anomalies using approaches like Linear Discriminant Analysis and Logistic Regression [1]. Researchers have employed various data mining strategies for real-time fraud detection by harnessing historical data [1]. Additionally, the use of the KNN algorithm and mechanisms for identifying outliers have been described as effective for detecting fraudulent actions [2]. The use of ensemble methods, including the Random Forest algorithm, has been explored to discern normal financial transactions and evaluate their efficacy against fraud detection methods employing neural networks [3].

For credit card fraud detection, one study [4] introduced an optimized Wale-algorithm for backpropagation to analyze transaction data. Further research [4][6] investigated previously classified data from imbalanced datasets to identify credit card fraud, utilizing K-means clustering to sample fraudulent transactions and applying genetic algorithms to group them. Various machine learning algorithms like KNN, Logistic Regression, and Naïve Bayes have been tested against available datasets, with studies indicating that KNN surpasses the other models in performance [2][6]. The evaluation of these methods was conducted using metrics such as precision, recall, Mathew correlation coefficient, and balanced classification rate specificity.

A novel fraud detection approach utilizing Big Data technologies is introduced in [7], which features the Scalable Real-time Fraud Finder (SCARFF) system utilizing analytical tools like Spark, Cassandra, and Kafka. This system boasts advantages in accuracy, fault tolerance, and the ability to scale. Finally, research in [8] presents feature engineering techniques aimed at reducing the rate of false positives, a common challenge in anomaly detection algorithms [16].

3. ML Methods for Financial Fraud Data Classification

3.1. Machine Learning tools and algorithm

A significant hurdle in fraud detection is its need for real-time processing [7]. Typically, the precision of manual fraud detection is suboptimal, making the process of spotting usual fraud patterns quite resource intensive. Compounding the issue is the dynamic nature of fraudulent behavior profiles, which are continually evolving. Additionally, the available data

on fraud is often biased and unreliable. The literature suggests a spectrum of Machine Learning (ML) methods to tackle these challenges, including general ML algorithms, ensemble methods, graph-based techniques, outlier detection, and deep learning. The success of automated fraud detection hinges on factors like the sampling method, variable selection, and the effectiveness of anomaly detection strategies [31]. Furthermore, the delay in reporting suspicious transactions exacerbates the problem, as fraud often goes undetected until customers raise the alarm [16]. Subsequent sections will delve into an in-depth exploration of anomaly detection and ensemble methods that aim to resolve these complications.

3.1.1 Outlier Detection Methods

An outlier is a data point significantly different from other data points in a dataset. Outliers can occur for various reasons, such as measurement errors, data entry errors, or natural variations in the data. They can significantly impact the outlier analysis in machine learning and interpretation of the data, so it is essential to detect them.

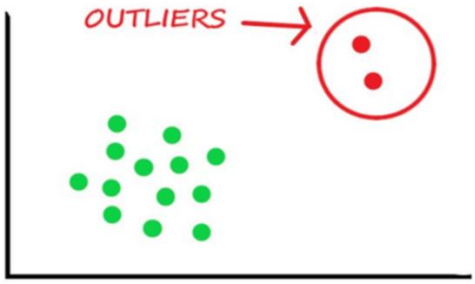


Figure 1. Outlier

Outliers can be detected using various methods, such as visual inspection of the data, statistical measures such as the Z-score or the interquartile range, or machine learning techniques. Once outliers are detected, they can be handled in various ways, such as removing them from the dataset, replacing them with the mean or median of the data, using outlier detection techniques using machine learning, or using algorithms that are less sensitive to outliers.

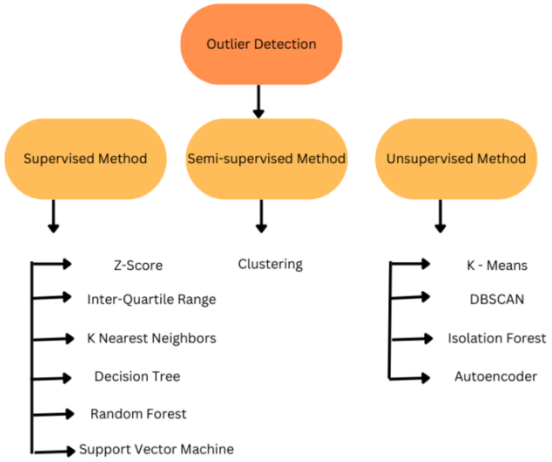


Figure 2. Outlier detection Methods

Local Outlier Factor: The Local Outlier Factor (LOF) method was initially proposed by Breunig et al. [83] for moderately high dimensional datasets. To reflect the degree of abnormality for the observation of an object locally (and not just globally on the whole dataset), a local outlier score (LOF) is calculated. The approach is local in the sense that only a restricted neighborhood is taken into account for the LOF score of each object. This is achieved by considering the k-nearest neighbors, a comparison of the local density of an object and the local density of its neighbor's objects. In the case that an object has a significantly lower density than its neighbors, it is considered to be an outlier. One of the very important parameters in this method is the value k, which needs to be set correctly. A too high value for k will detect just global outliers, whereas a low k results in the detection of outliers in small regions, which increases the overall false positive rate. In addition, a minor similarity exists between LOF and density-based clustering methods, such as OPTICS [84] or DBSCAN [85].

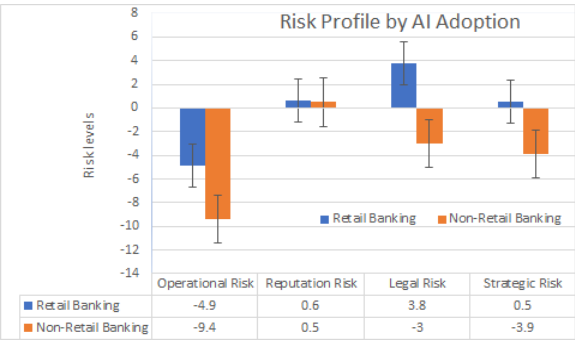
Isolation Forest: Another suitable method for outlier and novelty detection is Isolation Forest, which was proposed by Liu et al. [86]. This method is also suitable for high dimensional datasets. Isolation Forest does not—as many other methods—construct a profile or normal behavior. It isolates anomalies explicitly by relying on the fact that anomalies represent a minority in the dataset and that they have attribute values very different from normal ones. The isolation is performed in a tree structure, where anomalies that are closer to the root of the tree are isolated due to higher susceptibility than normal points. For this reason, the latter are isolated at the deeper end of the tree. After

building an ensemble of trees for a given dataset, anomalies are recognized by having a short average path length.

Elliptic Envelope: Another suitable method for outlier detection represents Elliptic Envelope [87]. Generally, this method is applied for Gaussian distributed regular data which in addition must not be high-dimensional. Under the assumption that the data are of Gaussian distribution, Elliptic Envelope fits an ellipse around the data with the help of robust covariance estimation. Any data point inside the ellipse is considered as inliers, whereas data points outside the ellipse are outliers. For fitting in an ellipse, a contamination parameter is used, determining the amount of data which will be inside the ellipse.

4. ROLE OF AI IN RISK MANAGEMENT AND GOVERNANCE

There are other potential benefits and opportunities provided by AI implementation. Consequently, there are challenges that need to be properly managed. Analysis shows that the main risks faced by retail banks by the quality of the data used for analysis, and the confidentiality of the data taken from the data store for analysis. No AI model can result from better accuracy unless the quality of the data considered for analysis is appropriate and reliable. To protect the privacy of the customer data a high level of confidentiality is to be maintained during the data analysis. Validation of the model uses is also another important requirement to achieve better performance and a high degree of interoperability to gain support from management and regulators. More adaptation of AI applications in banking operations leads to new challenges in the areas of operational, legal reputation, and strategies. The level of these risks is different for different types of banking services. Some of the retail banks believe that the implementation of AI may substitute human operators, and may add legal risks, but its impact on the reputation may be minimal.



	Operational Risk	Reputation Risk	Legal Risk	Strategic Risk
Retail Banking	-4.9	0.6	3.8	0.5
Non-Retail Banking	-9.4	0.5	-3	-3.9

Fig.8. Risk profile by AI adoption in banks

4.1 FRAUD ANALYSIS

Most banks adopt traditional rule-based methods of fraud analysis. Today due to the availability of advanced technologies the number of fraudsters is increasing, which is also an increased threat level to the banking industry. Fraud patterns are changing due to inconsistency in the banking systems. Fraud detection is possible with a valuable dataset

and a high-performance machine learning algorithm. The data are gathered from a public dataset and categorized, based on these we can classify the users as benign or fraudulent. Figure (9) gives the details about the fraud detection and prevention market size in 2016 – 2022, worldwide. Many statistical and machine learning models are used to analyze the fraudulent and non-fraudulent in each dataset. In this paper, we analyze popular statistical and machine-learning methods for the detection of a fraudulent transactions.

The most popular among these is Benford's law for modeling and the other machine learning modules for classification and binary decision trees [12]. These models help to determine benign and fraudulent transactions.

A. Benford's law : This law is used to determine patterns in a particular set of transactions or datasets[19]. Using this dataset can detect fraudulent transactions or anomalies. In Benford's law, the universal value of the data depends on the units. If there exists a universal probability distribution P(x), then it must be invariant under a change of scale as

$$P(k\,x) = f(k)P(x)$$

If $\int P(x)dx = 1 \quad then \int P(k\,x)dx = \frac{1}{k}$

Normalization denotes $f(k) = \frac{1}{k}$

Differentiating with respect to k and setting k = 1, we get

$xP^1(x) = -P(x)$ with the solution $P(x) = \frac{1}{x}$

This does not represent the proper probability distribution. The distribution of the first digit is shown as a percentage in figure (10). The frequency of the first digits follows the logarithmic relations as

$$Fa = log\left(\frac{a+1}{a}\right)$$

Fa is the frequency of the digit 'a' in the first place of used number. Table (1) gives the observed and computed frequencies.

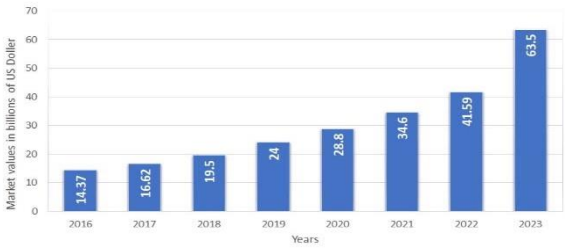


Figure 4: Fraud detection And Prevention Worldwide

The probability of the first digit D is given by a logarithmic distribution.

$$P_D = \frac{\int_D^{D+1} P(x)dx}{\int_1^{10} P(x)dx} = log_{10}\left(1+\frac{1}{D}\right)$$

For $D = 1, \ldots\ldots, 9$ For the second digit it is represented as

$$P_{D=D2} = \sum_{D1=1}^{9} log_{10}\left[1+\left(\frac{1}{D1\,D2}\right)\right]$$

D2 = 1, ..., 9 and so on When plugging in the digits 1 through 9, each subsequent digit has a diminishing probability that it will be the leading digit with 1 being the most common and 9 being the least.

Benford's law is widely used in executing accounting transactions and detecting fraud. Using Benford's curve income statement, general ledger, and inventory listing can be assessed and compared to the curve to determine its genuineness.

B. Machine learning classification algorithm

Under machine learning determining whether the transaction is fraudulent or benign is considered a classification problem. Different machine learning algorithms play a crucial role in fraud detection [21]. This includes Logistic regression, k- nearest neighbor algorithms, Random Forest (RF) Classifier, Support Vector Machine (SVM), and Naïve Bayes classifier. Among this algorithm it was found that Naïve Bayes classifier got the best accuracy. The comparative analysis of these classification algorithm is given in the figure (11).

C. Dataset

In this analysis of fraud detection, we used UCI dataset with balanced features like customer ID and the demographics details such as Customers' origin referring to zip code, type of the customer, age, gender, category, and amount of purchase when committed the crime. To avoid the imbalance in the dataset one can, perform oversampling or under sampling. We perform an exploratory data analysis on the dataset to capture its features. In data cleansing, the available categorical features are transformed into numerical values. An oversampled technique called SMOTE (Synthetic Minority Over-sampling Technique) is used here. It will create new data points from the minority class using the neighbor instances so that generated samples are not biased and the base accuracy score will improve.

Number interval	Observed frequency	Logarithmic interval
1 to 2	0.306	0.301
2 to 3	0.185	0.176
3 to 4	0124	0.125
4 to 5	0.094	0.097
5 to 6	0.080	0.079
6 to 7	0.064	0.067
7 to 8	0.051	0.058
8 to 9	0.049	0.051
9 to 10	0.047	0.046

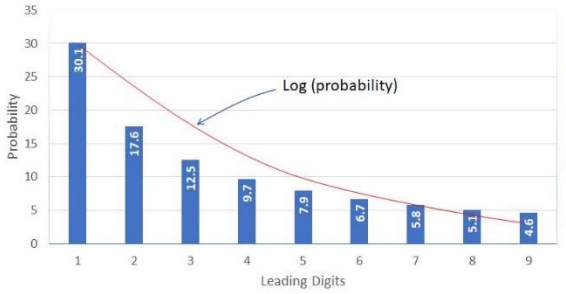


Fig.10. Observed and Computed Frequencies. Fraud detection using leading digits in Benford's law

Sl. No.	Algorithms	Accuracy
01	Logistic Regression	89.34%
02	K-nearest neighbors	93.45%
03	Support vector machine	94.89%
04	Decision Tree	96.81%
05	Random forest classifier	97.50%
06	Naïve Bayer's classifier	98.23%

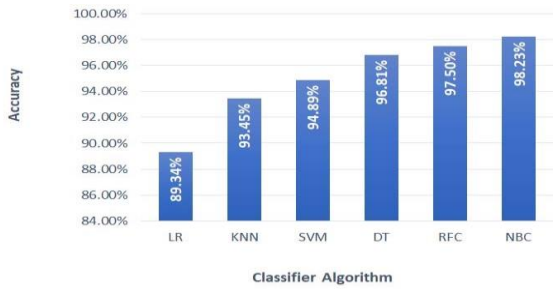


Fig.11. Classification Algorithms and their Accuracy

The KNN algorithm identifies similar things that exist in proximity[2]. The data points are closer to each other. The similarity can be calculated based on the distance functions such as Euclidean for the continuous variable as

The distance between the training data and the test data is obtained from the above equation, also then the k values related to the test data. Similarly, the distance between all the training cases with new value are calculated the in terms of distance.

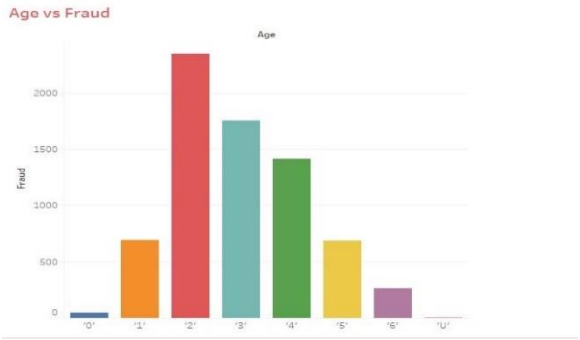


Fig. 12. Age versus Fraud

Random forest – RF is an ensemble classifier used in this both for classification and regression task. It involves the concept of bagging method, which is a collection of many weak learners. In RF they are considered as decision trees. One of the constraints of a decision tree is that they perform better for part of the dataset, but with a high variance due to greedy approaches of the model, because of this the approach may continuously select the best split at each level and it may not consider the current level. As a result of this there is a chance of overfitting. With this the performance of the model is better in training data and low in test data.

In RF this problem can be mitigated by using the bootstrapping method. In which the training data are trained randomly, where a different subsample of the data is used to train each decision tree. XGBoost – XGBoost is a gradient-

5. CONCLUSION

Use of machine learning algorithms proposed in this research to detect fraud in banking applications. The publicly available dataset from UCI is analyzed. The high level of imbalance in the dataset provided is highly biased toward most samples. This problem is tackled by the synthetic minority over-sampling technique (SMOTE). Implementation issues of this by KNN and Random Forest algorithms are handled by XGBoost as the boosting methods. The performance achieved using the model was 97.74%. In the analysis of the dataset, we found that people in the age group of 19-25 years

are more likely to be fraudulent than other customers' demography.

References)

[1] Baesens, B.; Van Vlasselaer, V.; Verbeke, W. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection; Wiley: New York, NY, USA, 2015. [Google Scholar]

[2] Zemankova, A. Artificial Intelligence in Audit and Accounting: Development, Current Trends, Opportunities and Threats-Literature Review. In Proceedings of the 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Athens, Greece, 8–10 December 2019; pp. 148–154. [Google Scholar]

[3] F. Black, "Toward a fully automated stock exchange Part I", Financial Analysts J., vol. 27, no. 4, pp. 28-35, Jul. 1971.

[4] F. Allen and D. Gale, "Stock-price manipulation", Rev. Financial Stud., vol. 5, no. 3, pp. 503-529, Jul. 1992

[5] Y. Cao, Y. Li, S. Coleman, A. Belatreche and T. M. McGinnity, "Detecting price manipulation in the financial market", Proc. IEEE Conf. Comput. Intell. Financial Eng. Econ. (CIFEr), pp. 77-84, Mar. 2014.

[6] D. Diaz, B. Theodoulidis and P. Sampaio, "Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices", Expert Syst. Appl., vol. 38, no. 10, pp. 12757-12771, Sep. 2011.