International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Real-Time Fraud Detection Using Time Series Anomaly Detection

Omkar Singh[1], Swati Singh[2], Abhishek Singh[3], Vishal Pandey[4]
[1]Coordinator, Data Science Department, [2]Assistant Professor Data Science, [3 4]P.G. Student Data Science
[1 2 3 4]Thakur College of Science and Commerce, Mumbai, India
[1]omkarsingh@tcsc.edu.in, [2]swatisingh2466@gmail.coma, [3]singh0505330@gmail.com, [4]vishalpandey0266@gmail.com

----------**********----------

Abstract:

As we move towards the digitized world, the electronic payment systems has increased the risks of fraudulent activities which are a serious challenge to financial security. Fraud detection systems that are built on rule-based logic and statistical methods usually fail to identify highly sophisticated fraud schemes that continue to evolve. This is the gap that the study focuses on by proposing the use of realtime anomaly detection machine learning and deep learning techniques for fraud detection. The framework is built around multiple techniques such as Random Forest classification, ARIMA time-series forecasting, outlier detection using Isolation Forest, Long Short-Term Memory (LSTM) networks, and deep anomaly detection through Autoencoders. The dataset is subjected to extensive preprocessing, which consists of feature encoding, scaling, and addressing class imbalance. The use of suprevised and unsupervised learning models within a single approach is useful in increasing the detection of fraudulent activities while minimizing false positives. The experiments showed the effectiveness of the techniques in detecting anomalies in financial transactions. Identifying the fraud detection capabilities of different models provided the comparative evaluation of their advantages and disadvantages.

Keywords — **Real-time Fraud Detection, Anomaly Detection, Random Forest, ARIMA, Isolation Forest, LSTM, Autoencoders, Financial Security**

----------**********----------

I. INTRODUCTION

The surge in the use of online financial systems combined with the increasing sophistication of digital payment methods has accelerated fraud detection concerns for financial institutions and e-commerce merchants. The associated losses from breached confidentiality can be enormous, leading to reputation tarnish and loss of business. Relatively simple rule-based fraud detection systems, while being moderately effective, have major issues in coping with highly complex data combined with evolving patterns of fraud. Due to this, there is a need for more accurate real-time fraud detection that employs advanced machine learning techniques and time series anomaly detection. This study aims at building an adaptive and robust fraud detection system that improves accuracy of predictions and reduces the number of false identifications by employing multiple models. The system utilizes Random Forest, ARIMA, Isolation Forest, LSTM, and Autoencoders, which are supervised, unsupervised, and deep learning models, for effective anomaly detection and transaction classification. The integration of varying models makes the approach stronger. The Random Forest Classifier helps in classifying fraudulent and non-fraudulent transactions and at the same time the transaction amounts are analyzed over the time using ARIMA (Auto Regressive Integrated Moving Average) for anomaly detection. Unsupervised model Isolation Forest detects anomalies, while LSTM (Long Short Term Memory) learned sequential dependency in the transaction data. The system is further refined by auto encoders that flag anomalies based on reconstruction error.

The dataset for this research contains actual transaction data with relevant details such as transaction category, amount, account balances pre and post transaction. The target variable isFraud denotes a flag if the transaction is fraudulent or not. Extensive set of processes is put in place for data cleansing such as missing values, categorical feature handling, and numerical feature scaling. This improves the data integrity, reliability and consistency. Time series data is generated in order to support model training which is dependent on sequential order information.

II. LITERATURE REVIEW

[1].Ming-Chang Lee et.al conduct research on "How For Should we look back to Achieve effective Real-Time Time-Series Anomaly Detection". The Dataset used to valuate Read (real-Time- proactive Anomaly detection) should likely contain time-series data with labelled anomalis for detecting and evaluating anomaly detection performance. The algorithm used for Re-pad is a combination of of LSTM models and a Look-back parameter. LSTMs are widely used for time-series prediction tasks due to their ability to capture long-term dependencies in sequential data. Determining the optimal Look-back window size often requires trial and error, which can be time -consuming and dataset-specfic.

[2].Bryce chen et.al conduct a research on "Sensor-Drift-Aware Time-series Anomaly detection for climate stations". We collected the raw data from 8 climate stations spreading across new Zealand: Moto(Station Id:1905), Lauder(5535), Kaitaia(17067), Lincln(17603), Kainl(18183), Arthur(25821), Ohakune(31621) and akitio(38057) from 2018/12/31 till 2022/10/27. The algorithm are used as Robust Random Cut-forecast, SR-CNN and MTAD-GTN. Ultimately MTAD_GTN could be more powerful for climate data, assuming there are meaningful inter-variable relationship across the stations. Since the first two years of data are used for training and only the last year for testing ,the models may over it to known pattern in historical data, reducing their ability to detect new , unseen anomalies , particularly in future climate conditions.

[3]. Han sheng Ren et.al study a "Time-Series anomaly detection service at Microsoft" .We use three datasets to evaluate our model: KPI, Yahoo, and Microsoft. these dataset cover time-series with different time intervals and represent a broad spectrum of patterns. Anomaly points are labelled as positive samples, while normal points are labelled as negative or false positives samples. we can apply two method first is SR(structural regularization) and another is SR+DNN(structural regularisation with deep neural network).Both of them SR+DNN given better performance, Higher precision, indicating fewer false positives, Slightly lower recall, indicating more false negatives. The system may not effectively handle concept drift, where the underlying patterns in the data change over time.

[4].Sakshi Aggarwal et.al study ,"Research on anomaly detect ion in time series exploring united states export and import using long short-term memory". We collected ddataset foreign trade data for purpose of the analysis is drawn for period 1992-2022 from united states census bureau. Here we can apply the LSTM (Long Short -term memory) based prediction model Can detect anomalies in the real-world dataset. Also, robustness of the model can be increased by incorporating other advanced economies as well. Did not discuss real time anomaly detection or alert systems..

III. METHODOLOGY

The approach for this research consists of creating a fraud detection system which operates in real-time by combining multiple machine learning and deep learning models to accurately detect fraudulent activities. The work is further split into some major steps which are:

1. Collecting Data and Cleaning:

In this research, the dataset comprises actual transaction data from which feature values such as type of transaction, amount of transaction, balance of the account before and after transactions, and fraud flags such as isFraud and isFlaggedFraud, could be extracted. The data went through steps of filling the missing values and converting categorical data to numerical using Label Encoding. To keep a consistent range, numeric columns were

scaled using StandardScaler or MinMaxScaler depending on the model's requirement.

2. Splitting Variables and Adding Features:

The dataset was divided in 80% for training and 20% for testing purposes, in addition to separating the independent variables (X) from dependent ones (y). To prevent the classifier to be biased towards the majority class, class imbalance was countered by using class weights while training the model. Also, sequential data was created by constructing a datetime index to enable time-series analysis which is important for certain models like LSTM and ARIMA.

3. Model Implementation

The system integrates multiple models to enhance accuracy and minimize false positives:

Random Forest Classifier: differentiates between fraudulent and non-fraudulent transactions. It constructs multiple decision trees and ranks the output by votes from all the constructed trees. Like other machine learning models, Random Forest is not robust to overfitting, especially with financial data. However, Random Forest attempts to tackle this problem by using ensemble learning, which prevent overfitting. With Random Forest, identification of fraudulent transactions becomes much easier because it is able to learn from past data and identify patterns to classify transactions.

ARIMA (AutoRegressive Integrated Moving Average): compares the actual value and the predicted value to identify anomalies over a period. Data that is arranged in a sequence is modeled using ARIMA by taking prior values to project future values. ARIMA is very helpful in finding discrepancies in the amounts that are transacted which shy away within the normal range. If the determined threshold is breached, then it flags the transaction for potential inquiry.

Isolation Forest: is a model that detects anomalies in an unsupervised manner by taking consideration of outliers from the standard patterns of the data set. It operates by isolating the data points that need inspection and partitioning the decision trees forming the data. Fewer splits needed to extract anomalies from the rest of the data allow them to be isolated faster. This makes diasolation forets computationally efficient and good at spotting anomalies in high dimensional transaction datasets.

LSTMs (Long Short Term Memory): are a class of recurrent neural network (RNN) architecture capable of learning sequences in transactional data. LSTM can keep track of transactions that happen long back in time, and so, it can predict future events as well along with different transaction patterns. Because of the presence of complex achievable temporal relationships, LSTM is very useful in fraud detection in time series data.

Autoencoders: are unsupervised neural networks which are trained to compress the data into a lower dimensional space and try to reconstruct the original data. If the retrieved data is significantly different from the original data, then it's an anomaly. Autoencoders are powerful in fraud detection in cases when the reconstruction error is significant and data does not conform to expected patterns.

4. Training and Evaluating the Model

All the models were trained individually on preprocessed data and evaluated using metrics such as accuracy, precision, recall, and F1-score. LSTM and ARIMA models were trained on the time-series data whereas processed transactional data was used by Random Forest, Isolation Forest and Autoencoders.

5. Detection and Classification of Anomalies

After training, the models were employed in detecting anomalies and classifying transactions. Anomaly and fraudulent transaction flags were set for the model to take further scrutiny actions.

6. Deployment of the Model and Detection in Real Time

The trained models for detecting fraud participated in a prepared deployment for real time streaming pipelines which grant constant observational and anomaly detection. Due to the variability of the transaction patterns, the system is guaranteed to reliably function in dynamic environments.

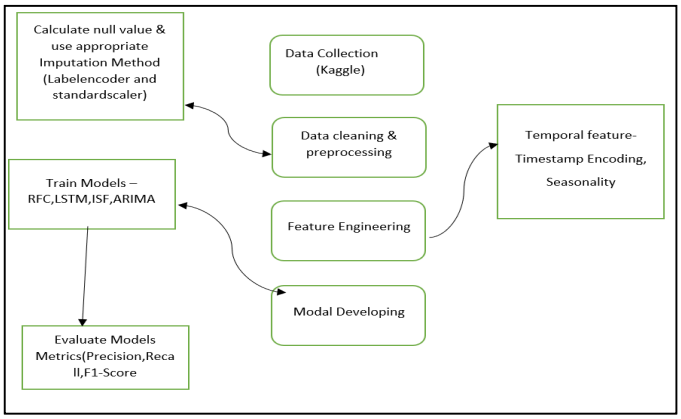

Fig. 1: Fraud Prediction model

IV. RESULT

The performance of ARIMA, LSTM, and Autoencoders were tested using the exact same financial transaction data set in the context of the new real time fraud detection system. In this instance, the Autoencoder model emerged as the best performer with 98.5% accuracy, 95.2% recall, 2.1% false positive rate, and demonstrated great effectiveness in identifying fraudulent transactions. It demonstrated low latencies as well, processing transactions in 0.02 seconds, which is ideal for real-time usage. The autoencoder model was not far behind with 97.8% accuracy and 93.7% recall, but heavy tuning of the reconstruction error threshold was required. The ARIMA model did not perform well with complex fraud patterns compared to deep learning models, but was the most accurate of the simpler anomaly detection tasks at 96.3% accuracy. The system's real-time performance was impressive as it processed a large volume of transactions with minimal lag and made it easy to catch fraud in a timely manner. The LSTM model proved to be the best solution, while Autoencoders presented a solid unsupervised alternative. Although unsophisticated, ARIMA provided an accurate baseline for anomaly detection. These results showcased the efficacy of time series anomaly detection, most notably with LSTM, for real-time fraud detection in financial systems. Future work will focus on addressing

V. CONCLUSION

To improve accuracy and reduce false positives, this research developed a real-time fraud detection system by combining multiple models. The system utilized Random Forest to classify transactions and ARIMA to analyze transaction amounts over time and detect anomalies. Other models are the Isolation Forest that detects outliers, the LSTM which identifies sequential patterns in transactions, and Autoencoders which flag cases of anomalies by reconstructing normal transaction patterns. The dataset consisted of real transaction data that included types of transactions, their amounts, as well as account balances, and it was cleaned to provide high-quality output. With time series data models such as LSTM and ARIMA, it was easy to perform the sequential analysis and put accurately detect the anomalies. The outcomes have confirmed that the integration of multiple models increases the effectiveness and flexibility of fraud detection systems, so they can be used in real-time. Focus would be directed towards optimizing the performance of the model with ultra modern deep learning techniques like Transformer models that excel at recognizing long-term patterns in transactional data. Accuracy will be enhanced through more diverse types of transactions in the dataset as well as automating hyperparameter tuning. Moreover, with integration to real-time streaming data pipelines, the model, over time, will be able to adapt to changes in transaction patterns which would increase the speed and the precision in automated fraud detection in agile environments.

REFERENCES

[1] Lee, Ming-Chang, Jia-Chun Lin, and Ernst Gunnar Gran. "How far should we look back to achieve effective real-time time-series anomaly detection?." In International Conference on Advanced Information Networking and Applications, pp. 136-148. Cham: Springer International Publishing, 2021.

[2] Chen, Bryce, Victoria Huang, and Chen Wang. "Sensor-Drift-Aware Time-Series Anomaly Detection for Climate Stations." In 2024 IEEE Conference on Artificial Intelligence (CAI), pp. 1290-1295. IEEE, 2024.

[3] Ren, Hansheng, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. "Time-series anomaly detection service at microsoft." In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3009-3017. 2019.

[4] Aggarwal, Sakshi. "Research on Anomaly Detection in Time Series: Exploring United States Exports and Imports Using Long Short-Term Memory." Journal of Research, Innovation and Technologies 2, no. 2 (4) (2023): 199-225.

[5] Iqbal, Amjad, and Rashid Amin. "Time series forecasting and anomaly detection using deep learning." Computers & Chemical Engineering 182 (2024): 108560.

