International Journal of Advance and Innovative Research
Volume 12, Issue 2 (XVII): April - June 2025
ISSN 2394 - 7780

ANOMALY DETECTION IN NEFT TRANSACTIONS

[1]Ayush Kumar Arun Kumar Mishra, [2]Sunita Jena and [3]Dr. Balkrishna Parab
[1]Department of Computer Science, Jai Hind College (Autonomous), Mumbai, India
[2]Assistant Professor, Department of IT, Jai Hind College (Autonomous), Mumbai, India
[3]Director, Aditya Institute of Management Studies and Research (AIMSR), Mumbai, India

ABSTRACT
Anomaly detection-a fundamental approach in data mining and machine learning-is based on finding any deviations in data patterns with respect to the expected behavior. This area of study finds application in various fields such as fraud detection, network security, and financial systems, where pinpointing any irregularities is key to safety and integrity. Anomaly detection in NEFT transactions was used to continually monitor the patterns of transactions for fraud or abnormality that would indicate security threats or human error. NEFT transactions represent high-volume data and require real-time access, so it is essential to use robust and powerful algorithms capable of finding minutiae in large datasets. In this study, three popular unsupervised anomaly detection algorithms-Local Outlier Factor (LOF), Isolation Forest, and Autoencoders-have been employed. LOF identifies various anomalous points according to deviations from the local density, thus effectively spotting outliers in transaction data interaction complexities. Isolation Forest does this by recursively isolating the anomalous data points, thus offering great efficiency in dealing with high-dimensional datasets. Autoencoders are a type of neural network with excellent data representation-learning ability; they mark a broad scope to reconstruct normal transaction patterns and flag all atypical patterns as anomalies. This research discusses and evaluates the relative performance of these methods on NEFT transactions, along with the salient features of each model in capturing anomalies within financial datasets. By investigating robust methodologies suitable for NEFT transactions, the study provides a significant boost towards enhancing fraud detection in financial systems.

Keywords *— Anomaly detection, NEFT transactions, fraud detection, Local Outlier Factor, Isolation Forest, Autoencoders, unsupervised learning, financialdatasets.*

I. INTRODUCTION
An unexpected occurrence or deviation, involving instances of data, events, or instances that differ significantly from the general distribution within data, anomaly detection becomes a favored substrate in the applications such as fraud detection, network security, and financial transaction monitoring for identifying irregularities that might indicate malignant activities or system failures. This work addresses the field of applying anomaly detection to NEFT transactions, hence within the mainstream of monitoring financial transactions, establishing identifying features of suspicious transactions or unforeseen patterns. Typically, datasets contain instances and attributes of machine learning, with instances representing groups of data pertaining to individual data points, while attributes signify certain features of the data in question. Datasets are either labeled or unlabeled depending on the information with respect to labels. Due to the presence of labels that explicitly indicate the class, supervised learning is trained on a labeled dataset, determining the mapping from inputs to specified outputs. In unsupervised learning, clustering and anomaly detection techniques accept input of any class without labels and hence find use when no prior knowledge of the class is available. In our case, NEFT transaction data is unlabeled; this allows us to run unsupervised algorithms that can independently identify anomalies without pre-existing knowledge of fraudulent labels [1].

A. Role of Anomaly Detection in NEFT Transactions
Anomaly detection processes in transaction activities of NEFT help to detect the emergence of irregularities for fraud detection, thus increasing operational efficiency, compliance with regulations, and customer protection. Anomaly detection helps identify unusual transaction patterns, such as unexpected transaction amounts or transaction frequencies, which might help wasting resources towards prioritizing high-risk cases. This will support AML and KYC requirements and helps assess customer risk by blocking unauthorized transactions. It also ensures the integrity of the system by pointing out potential problems thereby furthering the purpose of safeguarding the financial environment [2].

B. Significance of the Study
The effect of the study of anomaly detection in NEFT transactions is very crucial as it attempts to address some very relevant problems that modern financial systems face with regards to fraud detection, regulatory compliance, and client trust. With a fair understanding of the anomaly detection model, it will bring forth different tools for finding abnormal transaction patterns with greater accuracy. This will help to detect an early

ISSN 2394 - 7780

onset of abnormal transaction patterns and ultimately save resources and reduce the chances of financial crime, thereby protecting the users' assets. This study will also help financial institutions meet these sets of requirements for AML and KYC compliance, reducing the risk of regulatory penalties while improving their reputations. In addition, anomalous detection methods will enhance operational efficiency through process automation, thereby freeing resources to focus on other critical areas. In sum, this study enriches the electronic fund transfers' broader security and resilience, thus representing a further step toward establishing a safer financial environment [3].

II. METHODOLOGY

Several machine learning methods for anomaly detection rooted in the principles of LOF, Isolation Forest, and Autoencoders will be analyzed and looked at. Since the objective is to discover insidious and unexpected behaviors hidden within the unlabeled data, these three models provide complementary angles regarding how to understand anomalies.

- **Local Outlier Factor (LOF):** LOF is a density-based method that determines the local density deviation of a spatial pattern with respect to its neighboring points. LOF has been found suitable for outlier detection in close groups with irregular densities, enabling it to discover outliers based on local variations of the density [4].
- **Isolation Forest:** This ensemble model isolates anomalies through recursive partitioning of the data according to the feature values. Abnormal points tend to be isolated with a small number of partitions compared to normal data points. Isolation Forest is effective for high-dimensional data, and therefore fits its application suitability into a real large-scale anomaly detection scenario [5].
- **Autoencoders:** Autoencoders are a type of neural network partly trained to perform unsupervised learning. It works by training the network to produce a compressed representation of the given input data. The normal labeled data reconstructs well, providing a low reconstruction error, while anomalies would be poorly reconstructed relative to normal data. This approach is very good at modeling nonlinear relationships, which makes it effective in dealing with complicated data patterns [6].

III.DATASET REVIEW

The dataset of interest looks at transactions executed under NEFT. It consists of 33,249 entries and contains several factors pertinent to these transactions.

The month column encompasses 191 temporal entries during which the transactions were carried out. In the case of the financial institutions, 295 unique bank name entries are present with "B N Paribas" being cited as the many institutions. The dataset shows a significantly wide range for a number of transactions and amounts of money.

The no. of debit transactions column shows on average approximately 1,120,201 debit transactions with significant fluctuations around this mean as shown by the standard deviation. The maximum number of debit transactions recorded under the maximum quoted figure is 358,140,270.

The amt of debit transactions column describes on average a transaction amount of 749,987, though the standard deviation reflects quite some variation around the mean. The highest amount for debit transactions is 83,575,594.56. A similar trend is noticed in credit transactions. The no. of credit transactions column indicates an average of about 1,120,136 transactions, with variability suggested by meaningful standard deviation around this mean. The maximum recorded credit transactions total 193,142,132.

In terms of money, the amt of credit transactions mean is seen as 750,751 again with variable standard deviations indicating meaningful variability, and a peak of 73,590,501.40. To summarize, this database provides a picture of NEFT transactions across different banks and months, providing insights into the frequency and transaction value of these e-payments.

IV.IMPLEMENTATION

1. Data Preprocessing:

- **Dataset Loading:**

Dataset neft_transaction_metrics.csv is read by the function pd.readcsv which reads the NEFT transaction metrics into a pandas DataFrame [7].

- **Feature Selection:**

This involved selecting a subset of key features from the dataset:

ISSN 2394 - 7780

No of debit transactions, amt of debit transactions, No of credit transactions, amt of credit transactions, month, bank name. This selection emphasized debit and credit metrics, along with the transaction amounts, month, and bank name for better analytical insight.

- **Handling Missing Values:**

Identification consisted simply of numeric columns only for filling in missing values.

The NaN values of each column were filled with the mean to avoid bias and keep the central tendency of the data.

- **Data Scaling:**

In normalizing the dataset, only numeric features were selected. StandardScaler standardized the data to ensure that all numerical features had a mean of zero and a standard deviation of one. This enhanced the convergence rate and performance level of the model.

2. Exploratory Data Analysis (EDA):

To get good insights from the data at hand concerning transaction patterns and relationships between numerical variables, we undertook the following EDA processes:

- **Transaction Counts per Bank:**

By means of a count plot, we were able to visualize how transactions spread over different banks. This helped highlight which banks had more transaction volumes, contributing to understanding levels of engagement across banks.

- **Correlation Analysis of Numeric Columns:**

The correlation heatmap was created for the specified numeric columns covering transaction counts and amounts. This holds a consolidated view of how both transaction metrics relate to one another, pointing out potential dependencies or similarities.

Highly correlated variables indicate transactional patterns and help in feature selection for predictive modeling.

These visualizations provide foundational insights for identifying key patterns and relationships in the data, thus facilitating informed and critical data preprocessing and feature engineering decisions.

3. Clustering and Best Model Selection

Clustering methods, K-Means, Gaussian Mixture Model (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), were then applied to identify meaningful clusters from the dataset. Selecting the model that best accounts for the underlying structure in the data becomes the next stage of this research. In this case, the silhouette scores serve as the evaluation metric [8].

- **Model Initialization:**

Each clustering model is initialized with gear conducive to the dataset's characteristics.

a. **K-Means:** Set to create 3 clusters (n_clusters=3) with a fixed random state (random_state=42) for reproducibility.

b. **Gaussian Mixture Model (GMM):** Estimated 3 clusters (n_components=3) with the same random state for consistency in clustering behavior.

c. **DBSCAN:** Epsilon (eps=0.5) considers the distance between two points for them to be classified as neighbors. A minimum of 5 or more samples (min_samples=5) will mean that we can classify points as being in a dense region.

- **Model Fitting and Label Assignment:**

Each clustering model was fitted on the preprocessed data (features_scaled) to predict cluster labels. The predicted labels were stored as follows:

a. **kmeans_labels:** Cluster labels generated by the K-Means model.

b. **gmm_labels:** Linear model-generated cluster labels.

c. **dbscan_labels:** Cluster labels assigned by the DBSCAN model, where -1 indicates noisy points that do not belong to any cluster.

ISSN 2394 - 7780

- **Evaluation via Silhouette Score:**

To evaluate model performance, silhouette scores were computed for each model. The silhouette score computes how similar an object is to its cluster than to other clusters and a higher score is an indicator of better-defined clusters. **Silhouette score:**

K-Means: 0.9273

GMM: 0.4182

DBSCAN: 0.8562

Thus, based on the results, clustering observed that the K-Means model delivered model accuracy over the others as observed through the silhouette score.

- **Best Model Selection:**

Among the different clustering methods, the K-Means is dubbed, as per the silhouette scores, the best performing one.

4. Anomaly Detection Using the Best Model

For each of the clusters identified by the best-obtaining K-Means model, anomaly detection was run based on the following: Local Outlier Factor (LOF), Isolation Forest, and an Autoencoder-based neural network, building an ensemble approach for robust anomaly detection.

- **Initial Anomaly Flags:**

The results of each detection method were stored in three column flags added to the main data frame:

LOF-Local Outlier Factor Store.

Iso Forest-Isolation Forest Store.

Autoencoder-Autoencoder model store.

- **Anomaly Detection within Clusters:**

1. Local Outlier Factor (LOF):

In this case, a density-based method compares the local densities of data points with the lowest and highest outliers. It was configured with 20 neighbors (n_neighbors=20) and a contamination rate of 10%. Therefore, anomalies are flagged -1 and normal points 1.

2. Isolation Forest:

By way of this method, a tree-based model, the points are isolated by an artificial method through random splits. The method used is configured with a contamination rate of 10% with a fixed state. This means that it will also flag outliers as -1.

3. Autoencoders:

A neural network model was engineered with the objective of reconstructing the input data, thus labeling any point as an anomaly if it showed a significant reconstruction error. The Autoencoder architecture contains a chain of hidden layers with 32, 16, 8, 16, and 32 units. Each one was trained for 50 epochs, and using a cut-off (90th percentile) for significant reconstruction error, the model would flag such points as anomalies (-1).

- **Ensemble Anomaly Detection:**

For this reason, to take advantage of the qualities of each model in the procedure, an ensemble method was used whereby a new column, called ensemble anomaly, was created. It contained the count of the models detecting each point as an anomaly, thereby allowing a more trustworthy indication for anomalies based on the various detection methods.

- **Final Anomaly Assignment:**

A majority voting-protocol-based approach was used such that the final anomaly column flagged points that were detected as anomalies by at least two models. This threshold of two models helped increase the robustness of anomaly detection by making it mandatory that there is agreement between detection methods in order to reduce false positives and allow more reliable labeling of anomalies.

V. RESULTS

Three clustering algorithms, namely, K-Means, Gaussian Mixture Model (GMM), and DBSCAN were chosen in the current study to ascertain the optimum model for clustering the dataset. The models' performances were assessed using, and not limited to, the Silhouette Score.

ISSN 2394 - 7780

K-Means scored the highest Silhouette Score of **0.9273,** indicating well-defined clusters. DBSCAN was the second, scoring at **0.8562**, suggesting moderate performance in clustering. Finally, GMM scored the lowest with a score of **0.4182**, indicating less distinguishable clustering. Hence, K- Means model based on this performance as the tightest clustering model was chosen to be the best clustering predictor for this data The NEFT systems trans actual analysis on the measure of transaction amounts reveals point anomalies as:

- **The number of Credit Transactions:** The high proportion of anomalies is observed on the lower range transaction amounts indicating that some monthly outlier spikes are apparent and unusual.
- **The Number of Credit Transactions:** Monthly outlier spikes in the frequency of transaction occurrence, implies that anomalies are limited to the lower boundary of transactions.
- **The amount of Debit Transactions:** The largest dispersion of anomalies is recorded with regards to this metric with very high and low anomalous values occurring suggesting a strong time-based seasonality in debit amounts fluctuating through time.
- **The number of Debit transactions:** Temporal Debit recurring spikes are well stabilized as the transactions count rises irrespective transaction amounts, with the exceptions of lesser occurrences indicating stronger anomalies.

Overall Finding: The **amount of Debit transactions** which are notable under this metric emerge with the widest spread, suggesting that these types of transactions have a less uniform occurrence and can be more used to measure the presence of outlier activity [9].

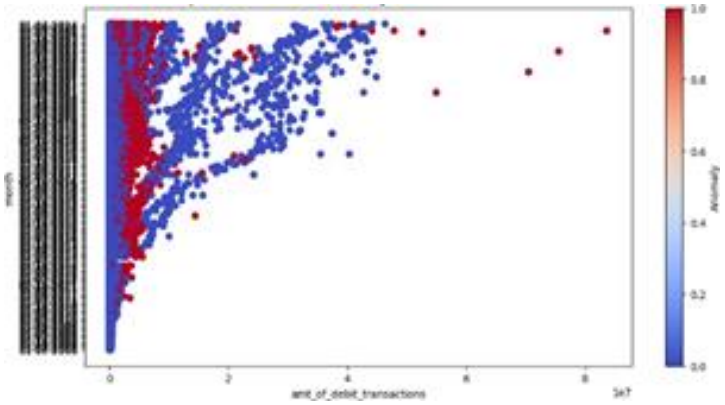

Fig.1 Anomaly Detection with Best Clustering Model and Ensemble Methods

VI. CONCLUSION

This study uses a comprehensive clustering and error detection framework, comparing three clustering models—K-Means, Gaussian Mixture Model (GMM), and Spatial Clustering with Noise (DBSCAN)—to identify the best model. Based on the silhouette scores, K-Means emerged as the best model to capture well-defined clusters in the dataset.

A combination of three methods is used to detect anomalies within each identified group: local outlier factor (LOF), isolation forest, and autoencoder-based neural networks. Each method has unique advantages – LOF for local density sensitivity, Exclusion Forest for efficient outlier isolation, and Autoencoders for reconstruction-based anomaly detection. Errors are scored if detected by at least two of these methods, providing a standardized approach that minimizes false positives.

By combining clustering and multi-model anomaly detection, this approach provides a powerful framework for understanding data structure and identifying true outliers within each cluster. This compromise approach ensures greater reliability and flexibility for complex datasets, enabling accurate collection and sensitive anomaly detection for future analysis.

REFERENCES

1. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323.
2. Al-Doghman, J. S. S. E. B. (2011). Anomaly detection techniques for fraud detection. Journal of Financial Crime, 18(2), 97–103.
3. Hasan, M. H. Z., Ganaie, M. A., & Khan, M. M. U. (2014). Anomaly detection in financial transactions for fraud detection. Computers & Security, 45, 152–169.

ISSN 2394 - 7780

4. Breunig, M. R., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (pp. 93–104).
5. Liu, L., Wu, J., & Wu, Z. (2011). Isolation forest. In 2011 IEEE 11th International Conference on Data Mining (pp. 413–422).
6. Bengio, Y. (2009). Learning deep architectures for AI. Foundations and Trends® in Machine Learning, 2(1), 1–127.
7. The pandas development team. (2025). Pandas: A powerful data analysis toolkit. Retrieved April 2025, from https://pandas.pydata.org
8. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining(pp. 226–231).
9. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 (pp. 281–297).