Topological Data Analysis to Aid Machine Learning Models in Detecting Fraudulent Bank Transactions

Taiwo Onitiju
School of Computing
University of North Florida
Jacksonville, Florida, USA
N01578746@unf.edu

***Abstract*—Financial fraud detection stays a vital challenge for financial institutions due to the growing complexity of fraudulent pursuits by fraudsters as well as the inherent class imbalance of fraud to legitimate transactions in datasets. This paper investigates the use of Topological Data Analysis (TDA) - a field that implements topology, a branch of mathematics for studying data shape in order to improve machine learning models in identifying fraudulent transactions.**

I offer a hybrid approach that incorporates persistent homology features extracted from TDA with conventional supervised learning techniques (i.e., XGBoost, Random Forest, and Neural Networks). My experiment on using synthetically generated bank transactions of about 1,000,000 entries demonstrated that while TDA provided interpretable structural insights, its computational complexity does limit performance gains compared to conventional feature engineering. [1].

***Index Terms*—Topological Data Analysis, Persistent Homology, XGBoost, Imbalanced Learning**

I. Introduction:

A. Problem Statement:

An evergrowing concern in the present-day world are scams, particularly financial scams which appear to trend alongside the rise of online financial transactions and operations. Due to this, the need to have up-to-date systems and techniques for detecting these scams is crucial for protecting companies and individuals from financial loss, keeping trust and faith in the global financial system, and dissuading more individuals and groups from viewing scamming as an easy avenue for profit gain. Conventional techniques of scam detection often rely on statistical techniques in combination with machine learning models such as:

- Rule-based heuristics (easily bypassed).
- Supervised machine learning models (limited by class imbalance).
- Deep Learning (requiring extensive labeled data).

While they have been proven effective to an extent, many of these approaches often overlook complex patterns hidden within the data that would prove crucial in detecting such fraud.

B. Research Goal:

Topological Data Analysis (TDA) is an emerging sector in data science that brings a unique way to see such complex patterns by analyzing the shape of the data [1]. TDA offers a collection of techniques that revolves around the idea of data having a shape that can visualize relationships other methods may miss [2]. Recent applications in fraud detection [3], [4] demonstrate its potential to uncover sophisticated financial crimes that evade traditional machine learning methods [5]. My research evaluates whether TDA extracted topological features can enhance machine learning models by capturing higher dimensional transaction patterns that conventional method might miss.

The main goal of this study is not to revolutionize financial security but rather to investigate the effectiveness of TDA in improving financial scam detection by taking advantage of its capability to analyze intricate forms within financial datasets.So thereby, utilizing TDA, this study will contribute to a better understanding of fraud patterns while also improving detection processes in an effective way.

II. Background/Literature Review:

A. Topological Data Analysis:

Statistical models such as logistic regression and neural networks that are being used, while effective often over-rely on feature engineering where specialists need to specify important characteristics for detecting fraud. This can be restricting especially when complex fraud patterns come into play. In addition, such models are susceptible to hostile attacks, in which scammers adapt their behavior to prevent detection. These systems also face challenges with class imbalance [6] in addition to the evolving attack patterns [7]. My proposed hybrid approach involves combining TDA with supervised learning [8] as it show promise in addressing these limitations.

Topological Data Analysis is a rising technique stemming from the mathematical sector that aims to address some of the limitations discussed earlier. TDA studies the shape of data to answer questions through hidden patterns analysis. This approach is already being applied in different areas such as medicine, genomics, and natural language processing among many others, but its use in financial fraud detection is a fairly new approach with the potential to offer useful insight that would aid in effectively combating fraud attacks [2].

Compared to past research, there has rarely been a use for implementing TDA techniques to improve the detection of financial scams. Since TDA focuses on the shape and link between transactions in a manner that would aid machine

learning models, it becomes a new viewpoint through which transactional data could be analyzed. This study's purpose is to make use of TDA to locate complex fraudulent patterns, stacking another layer/check that conventional detections methods may not catch. TDA applies two major tools from algebraic topology to analyze data shape:

TDA Technique	Description	Relevance to Fraud
Persistent Homology	Tracks multiscale features (H0=clusters, H1=loops)	Detects fraud rings (H0) and laundering cycles (H1)
Mapper Algorithm	Constructs a graph summary of data	Visualizes transaction networks

The mathematical framework of persistent homology [2] enables a wide scale analysis of data shape, while the Mapper algorithm [9] provides graph-based representations that can uncover hidden information through the patterns formed.

B. TDA Related Works:

The implementation of TDA in relation to fraud and scams, while not widely popular, has been investigated in different areas such as:

- [4] Fan et al. in a paper published in 2022, used TDA for Ethereum scam detection (AUC-ROC: 0.91).
- [10] Monkam et al in a paper published in 2023, applied TDA to network intrusion, showing 12% recall improvement.
- [11] Grover et al in a paper published in 2022, highlighted class imbalance challenges in fraud datasets.

III. Methodology:

The implementation of my hybrid TDA-ML pipeline was made up of four key stages:

- Data Preprocessing & Feature Engineering
- Topological Feature Extraction
- Model Training & Evaluation
- Computational Optimization

A. Data Preprocessing & Feature Engineering

1) Dataset: Synthetically generated bank account fraud datasets retrieved from Kaggle **(Bank Account Fraud Dataset Suite (NeurIPS 2022)** were used containing:

- 1 million transactions (30 features, 3% fraud rate)
- Key features:
  - Temporal: days-since-request, month
  - Behavioral:velocity-6h, velocity-24h
  - Risk indicators: credit-risk-score, device-fraud-count

2) Preprocessing Steps: Missing Data Handling: I handle missing values using robust imputation techniques, addressing the high missingness rates noted in financial datasets [11].

- Numeric: Median imputation for prev-address-months-count (contained over 712,000 missing values).
- Categorical: "-1" for addressing "Missing" categories (for example payment-type).
- Novelty: I incorporated a binary missingness indicator feature in order to use the "missingness" as a potential link to detecting fraud (i.e., credit-risk-score-missing).

Feature Scaling:

- Min-Max normalization for bounded features such as the income.
- Robust scaling for heavy-tailed features such as the velocity-6h.

Class Imbalance Mitigation:

- SMOTE-ENN (Synthetic Minority Oversampling + Edited Nearest Neighbors) resampling for extreme imbalance cases (under 5%).
- Also Implemented other sampling techniques such as ADASYN and SMOTE depending on the ratio of imbalance.

B. Topological Feature Extraction:

1) Persistent Homology Pipeline: My Vietoris-Rips implementation builds on optimized algorithms, with computational enhancements for large-scale banking data. To generate the persistence homology (also called the persistence diagram), the appropriate distance matrix computation had to be selected:

- Tested metrics: Euclidean, cosine, Manhattan
- Optimal choice: Euclidean/Cosine as they both preserved both the local/global structures best among all metrics.

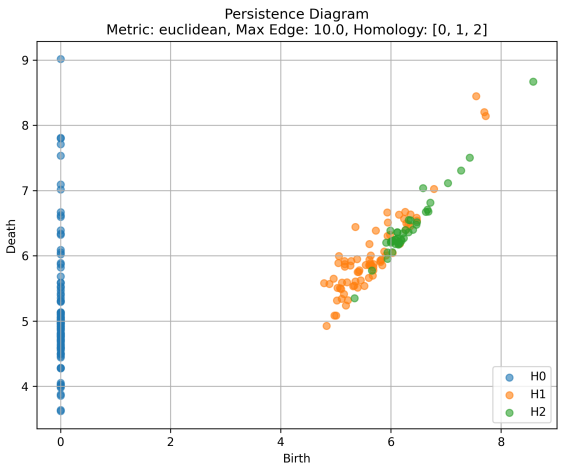

Fig. 1. Euclidean Persistence Diagram Visualization.

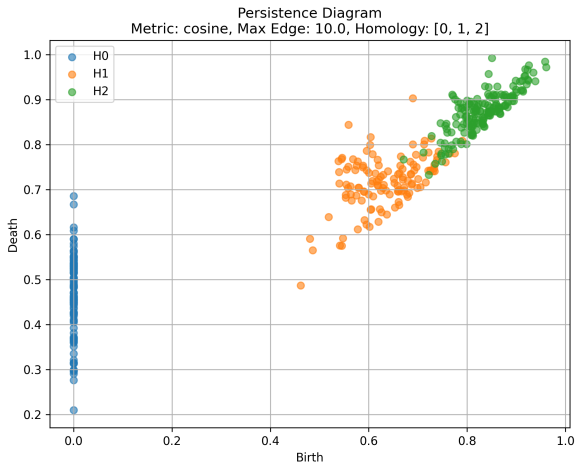

Fig. 2. Cosine Persistence Diagram Visualization.

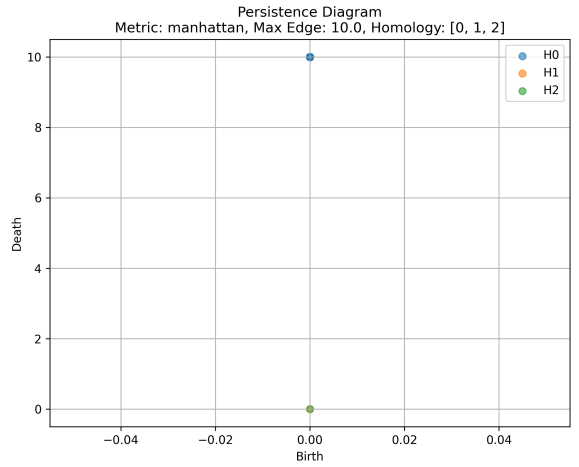

Fig. 3. Manhattan Persistence Diagram Visualization.

Generation of the persistence diagram involved using the Vietoris-Rips Complex (which is is a technique of forming a topological space from distances in a set of points) which was available as a python package.

```
def compute_persistence_diagrams(X, max_edge_length=np.inf, homology_dimensions=[0, 1]):
        print(f"Auto-computed max edge length: {max_edge_length:.2f}")

    VR = VietorisRipsPersistence(
        metric="euclidean",
        homology_dimensions=homology_dimensions,
        max_edge_length=max_edge_length,
        collapse_edges=True,
        n_jobs=-1
    )
    return VR.fit_transform(X)
```

Fig. 4. Vietoris-Rips Complex Computation.

- Batch processing: As mentioned earlier, TDA is computationally expensive, as such, to by address this issue, when working with extremely large datasets, I implemented mini-batch processing of 10K samples to prevent memory resource related errors.

Feature Extraction From Diagram:

- Persistence Images: 20×20 vectorized diragram
- Persistence Statistics:
  - Mean lifetime (H0 and H1)
  - Entropy (H0 and H1)
- Betti numbers: Counts of H0/H1 features present)

Fig. 5. Mapper Graph Shape Of 20,000 Transactions.

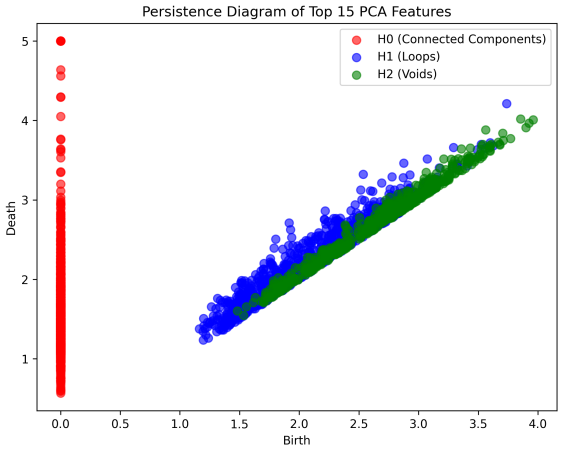

Fig. 6. Persistence Diagram Of 20,000 Transactions.

C. Model Training:

1) Feature Integration Approach: Training the model involved combining the original dataset features + with the engineered features + the TDA persistence features (images, statistics, and entropy) extracted from the persistence diagram. Dimension reduction was also applied post-concatenation in order to lower the use of redundant data.

2) Model Configuration: For each model trained, I implemented hyperparameter tuning to determine the best configuration for each model to attain the most accurate result. Grid search was implemented for both XGBoost and Random Forest, while Keras tuner was used for Neural Network.

Model	Key Hyperparameter	TDA Handling
XGBoost	max-depth = 6, scale-pos-weight = 33	Feature importance pruning
Random Forest	n-estimators = 500, class-weight = ”balanced”	Permutation importance analysis
Neural Network	layers = [64,32], dropout = 0.3	Leanred embedding layer

D. Computational Optimizations:

To more efficiently address the high computational cost, I implemented parallelization in the resampling and hyperparameter tuning for model training. Dividing the dataset being ran into batches of smaller sub-datasets that can be executed simultaneously across multiple processors proved effective in allowing my system to better handle the high load demand.

IV. Results & Analysis:

1) Topological Insights:

- Persistence Diagram: From fig. [6], the persistence diagrams revealed a clear grouping of H0 points (clusters of connected components) displaying a fast birth to death cycle. H1 (loops) and H2 (voids) display patterns that could match fraud aligning with findings in [10] .

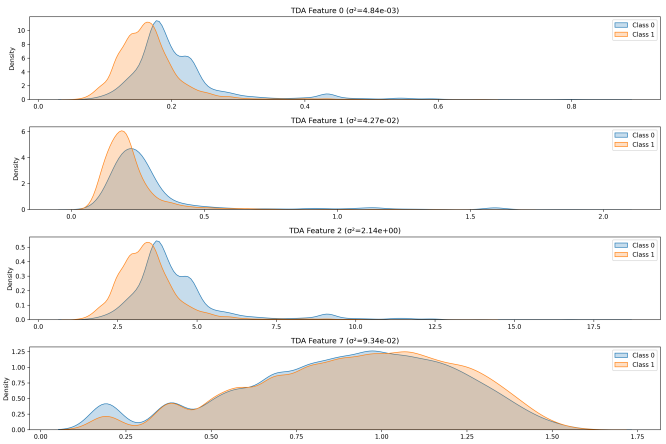

Fig. 7. TDA Feature Extraction Visualization From Persistence Diagram.

- Mapper Graph: From fig. [5], I noticed many nodes highlighted potential for fraud (marked as purple). This is due to the extreme imbalance of the dataset, resulting in the mapper graph being unable to effectively distinguish being fraudulent and non-fraudulent transactions unless resample occurs to balance the ratio.

Fig. 8. Mapper Graph with resampling to achieve 50% - 50% Ratio.

2) Performance Metrics:

Model	AUC-ROC (TDA)	Precision	Recall	F1-Score
XGBoost	0.900,	45.22%,	34.67%,	0.3925
Random Forest	0.891,	39.35%,	34.67%,	0.385
Neural Network	0.856,	42.56%,	72.00%,	0.320

The above table shows the performance of dataset with the following parameters:

- Dataset Size: 20,000 transactions.
- Fraud - Non Fraud Ratio: 5% - 95%

It should be noted that the TDA features contributed significantly versus the original features and the engineering features, offering a general improved precision (less false positives) but lowered recall (fraud detection rate).

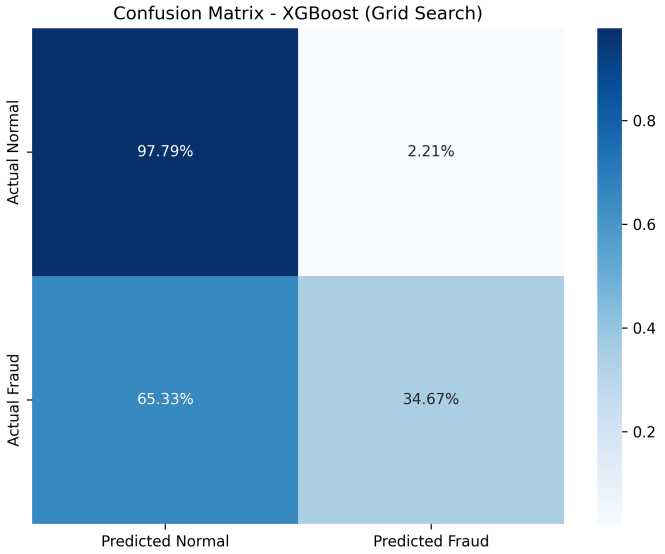

Fig. 9. XGBoost Confusion Matrix.

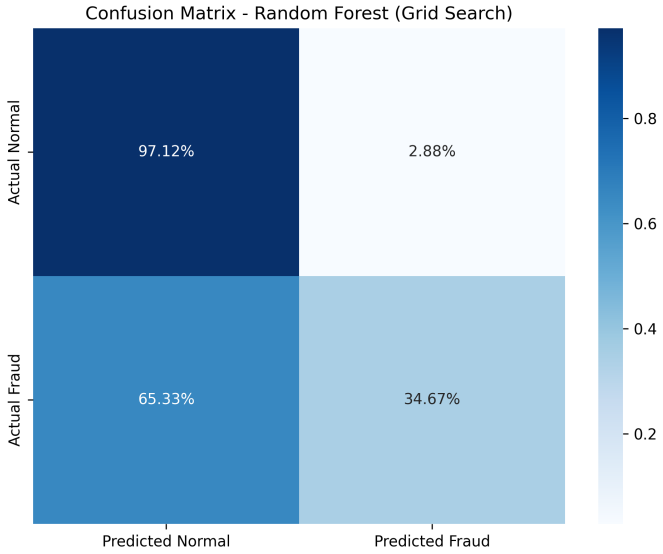

Fig. 10. Random Forest Confusion Matrix.

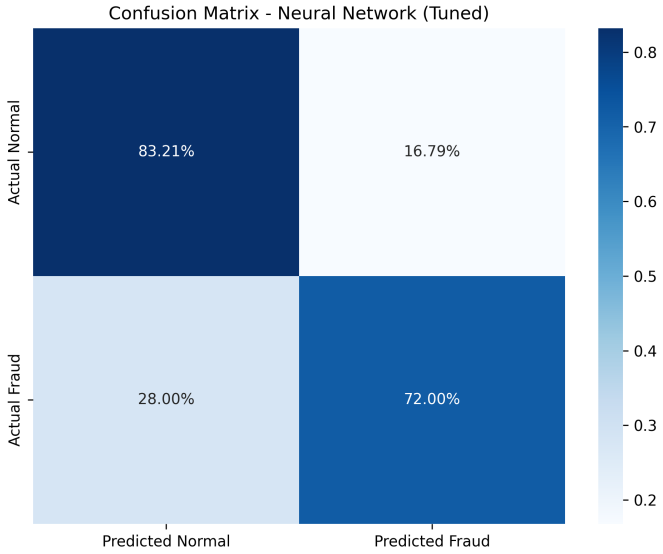

Fig. 11. Neural Network Confusion Matrix.

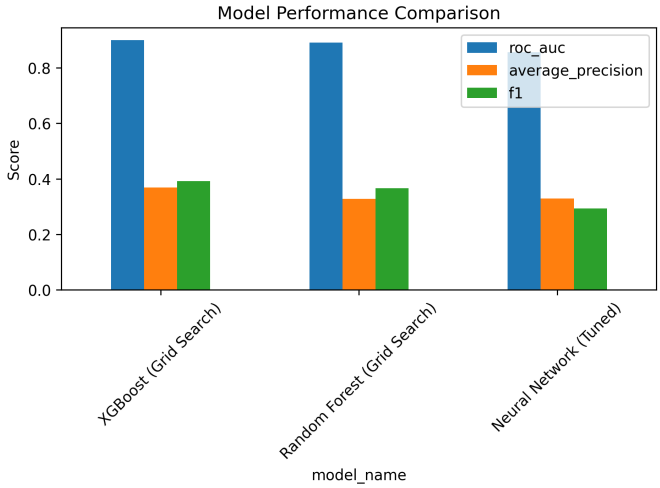

Fig. 12. Model Comparison.

V. Ethical Considerations:

Since the dataset used where synthetically generated, this project provided no ethical risks. However, making use of actual transaction data offers serveral factors that will need to be taken into account. Information present such as personal names, account numbers, etc. must be carefully handled. as the dataset must maintain anonymity by omitting or modifying said data to ensure individuals and organizations are not put at risk while also ensuring the data remains accurate for use in testing and analysis [12]. Any restriction or licensing agreements imposed by the data provider will need to be adhered to and the dataset will be used in accordance with it [13]. Clear information on how the data is used will also need to be laid out in the research from data retrieval, to analysis so as to ensure transparency and accountability for future researchers to build upon the work ethically.

VI. Conclusion:

This study investigated the use of Topological Data Analysis (TDA) to improve machine learning models for bank fraud detection. While TDA provided explainable structural insights—particularly in identifying isolated fraud clusters (H0 components) [14], its practical utility was limited by computational costs (which mirrors challenges reported in [**?**]) and marginal performance gains compared to traditional feature engineering.

The hybrid TDA-ML approach achieved decent recall in detecting organized fraud rings, showing potential for specialized use cases, but required significantly higher training time and additional labeled data for stable results. Notably, higher-diemensional topological features (H1 loops) showed limited predictive value, suggesting that traditional fraud patterns may not show the cyclic structures TDA excels at capturing.

For future endeavors, work should explore real-time implementations [**?**] and hybrid graph-based architectures [8] to close the gap between topological insight and operational

scalability. While not yet a universal solution, TDA gives a mathematically rigorous framework for detecting complex financial crimes that evade traditional detection methods. .

In conclusion, implementing TDA provides a favorable avenue in tackling increasingly complex scams. By highlighting the interaction between machine learning models and TDA, the importance of multi-layered detection methods with the potential to attune with evolving threat tactics is offered.

REFERENCES

[1] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[2] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete & Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

[3] S. Dey and M. Bhatia, "Anomaly detection in financial transactions using topological data analysis," *International Journal of Information Management*, vol. 57, p. 102131, 2021.

[4] S. Fan, S. Fu, Y. Luo, H. Xu, X. Zhang, and M. Xu, "Smart contract scams detection with topological data analysis on account interaction," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 468–477.

[5] M. Fitting, "Financial fraud detection using machine learning: A review," *International Journal of Computer Applications*, vol. 159, no. 6, pp. 5–10, 2017.

[6] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *IEEE Symposium on Computational Intelligence and Data Mining*, 2015, pp. 159–166.

[7] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 1–21, 2016.

[8] F. Hensel, M. Moor, and B. Rieck, "A survey of topological machine learning methods," *Frontiers in Artificial Intelligence*, vol. 4, p. 681108, 2021.

[9] F. Chazal and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientists," *Frontiers in Artificial Intelligence*, vol. 4, p. 667963, 2021.

[10] G. F. Monkam, M. J. D. Lucia, and N. D. Bastian, "Preprocessing network traffic using topological data analysis for data poisoning detection," in *IEEE Conference on Dependable and Secure Computing*, 2023, pp. 1–8.

[11] P. Grover, J. Xu, J. Tittelfitz, A. Cheng, Z. Li, J. Zablocki, J. Liu, and H. Zhou, "Fraud dataset benchmark and applications," *arXiv preprint*, 2022.

[12] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[13] B. J. Zikmund-Fisher, A. Fagerlin, and P. A. Ubel, "Ethical considerations in the use of data for research," *Health Affairs*, vol. 29, no. 3, pp. 540–546, 2010.

[14] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[15] C. C. Albrecht and K. Garrison, "An overview of fraud detection: Techniques and applications," *Journal of Economic Perspectives*, vol. 32, no. 2, pp. 213–224, 2018.

[16] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, 2011.

[17] J. West and M. Bhattacharya, "Intelligent financial fraud detection practices," *Electronic Commerce Research*, vol. 16, no. 3, pp. 373–409, 2016.

[18] H.-P. Kriegel, P. Kröger, and A. Zimek, "Geometric approaches to anomaly detection," *Statistical Analysis and Data Mining*, vol. 4, no. 1, pp. 31–39, 2011.

[19] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[20] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[21] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.

APPENDIX

A. Code Repository:

The link redirects to the GitHub repository for where the code, associated files and model results are stored: GitHub Repository

B. Supplementary Figures

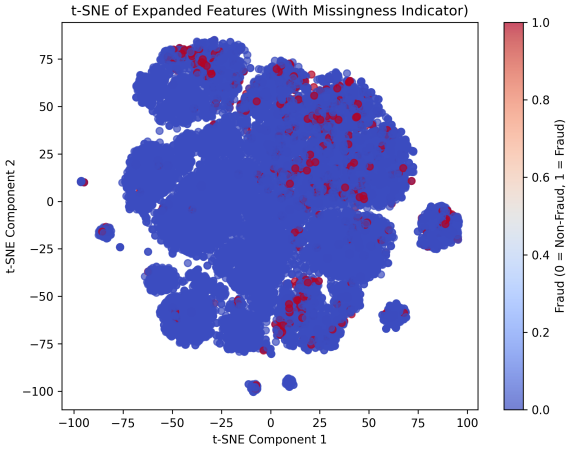

Fig. 15. TSNE expanded features plot.

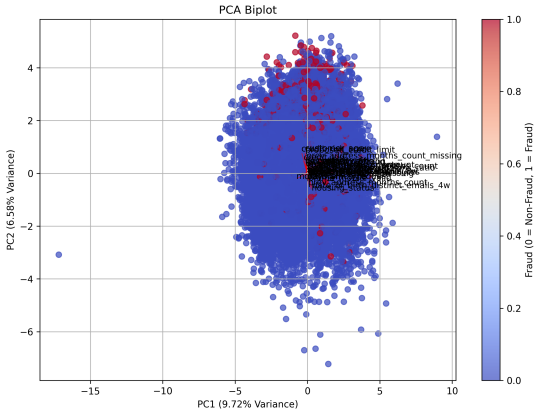

Fig. 13. PCA-biplot.

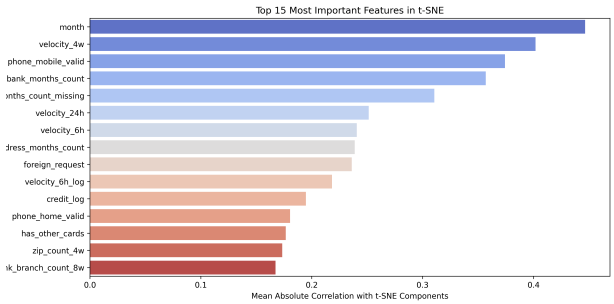

Fig. 16. TSNE Feature Importance.

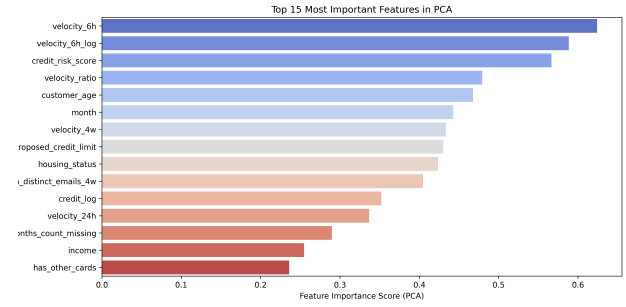

Fig. 14. PCA Feature Importance.

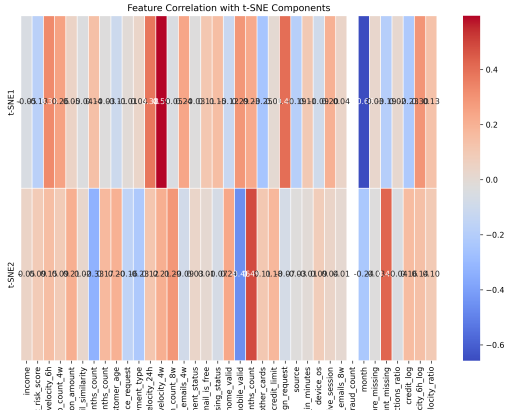

Fig. 17. TSNE Feature Correlation.

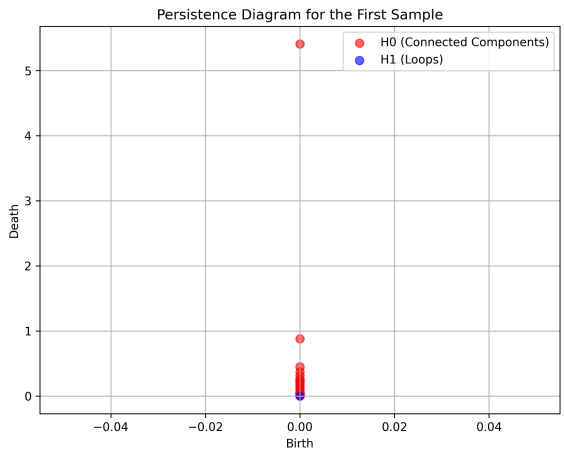

Fig. 18. Persistence Diagram of Single Sample.

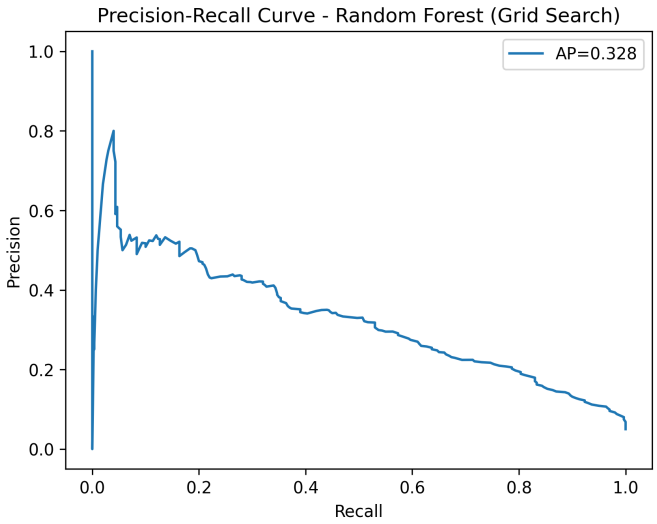

Fig. 20. Precision-Recall Random Forest.

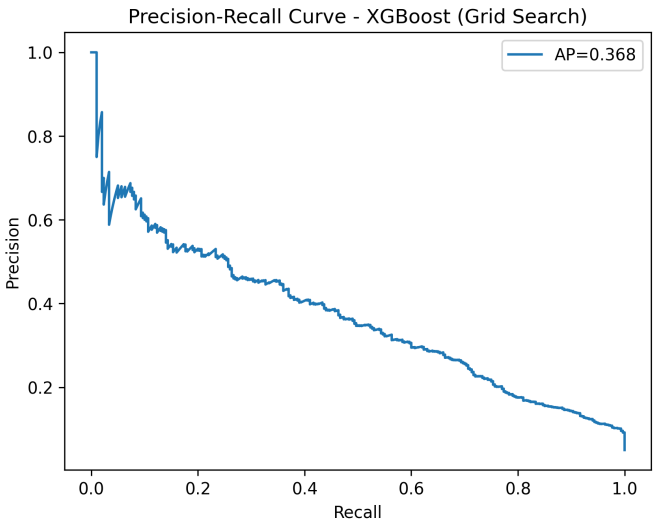

Fig. 19. Precision-Recall XGBoost.

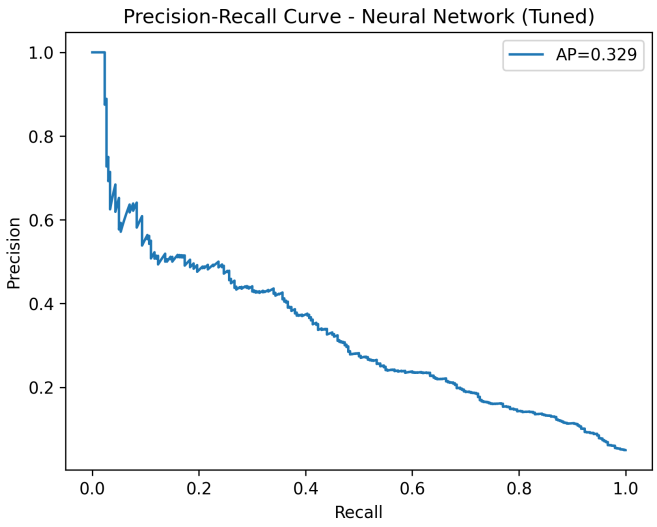

Fig. 21. Precision-Recall Neural Network.

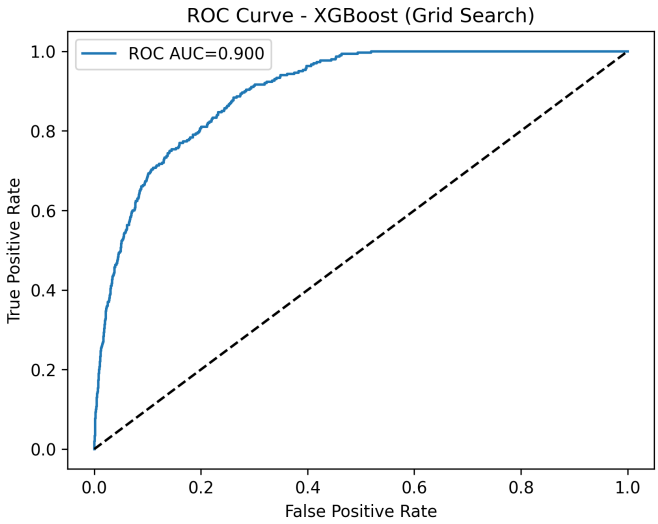

Fig. 22. ROC Curve XGBoost.

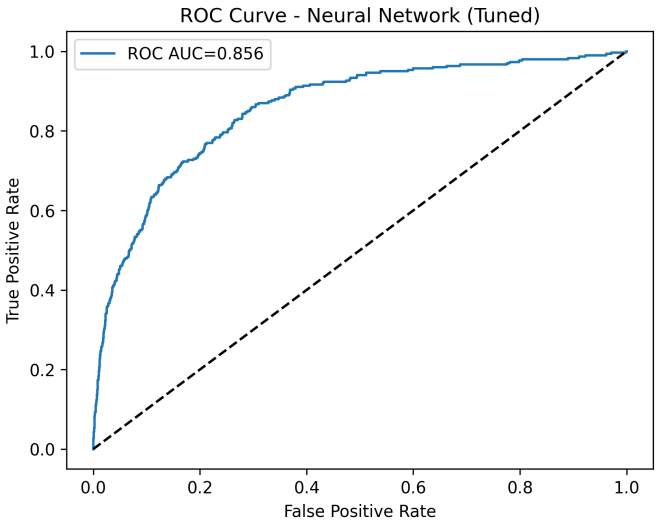

Fig. 24. ROC Curve Neural Network.

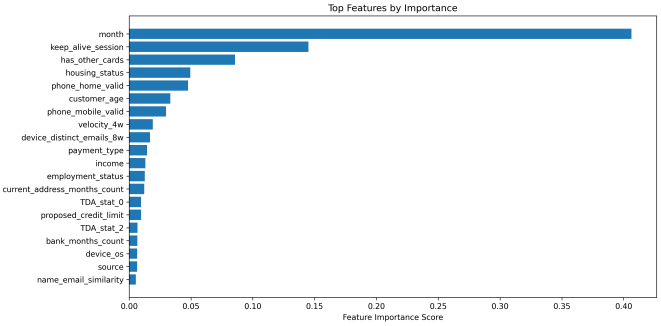

Fig. 25. Feature Importance Post TDA

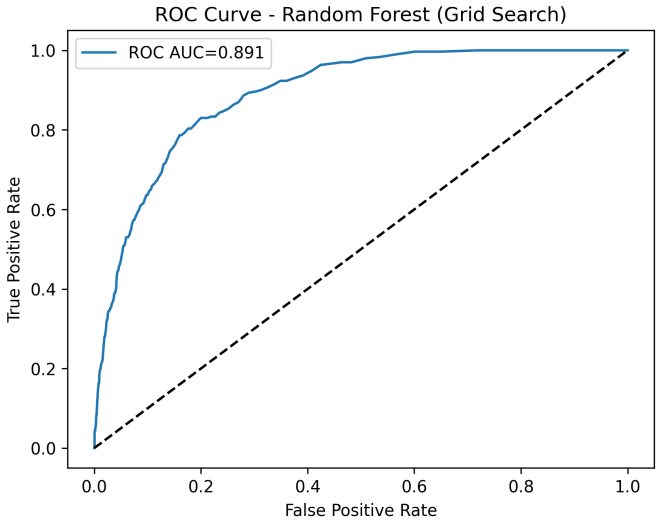

Fig. 23. ROC Curve Random Forest.