Explainable and Fairness Driven Approach to Fraudulent Bank Account Detection

Nithika Radhakrishnan
Pennsylvania State University
University Park, USA
nkr5274@psu.edu

Jay Patel
Pennsylvania State University
University Park, USA
jpp5955@psu.edu

Apurva Sista
Pennsylvania State University
University Park, USA
avs7350@psu.edu

Mohamed Elmanzalawi
Pennsylvania State University
University Park, USA
mye5156@psu.edu

Figure 1: Symbolizing Fraudulent Account Detection

Abstract

In today's financial sector, fraud detection methods that depend on machine learning are viewed as unreliable due to their hidden or black-box nature. As a result, financial institutions continue to rely on conventional practices such as flagging abnormal account activities and unusual account holder information to detect fraud. While these detection methods accurately identify known fraud patterns, they are more susceptible to complex and relatively new fraud tactics. The constant emergence of new fraud practices creates an urgency to capitalize on the unique pattern recognition capabilities of machine learning models, especially when the proper steps are employed to verify and explain their predictions. The following analysis provides a practical example of an accurate and transparent machine learning approach to detecting fraudulent bank accounts.

Keywords

Fraud Detection, Bank Account Fraud (BAF), Binary Classification, Model Explainability (SHAP), Machine Learning Fairness.

1 Introduction

1.1 Background

Bank fraud refers to the use of illicit methods to deceive and steal from financial institutions. It has taken many forms throughout the 21st century; the most common types of Bank Fraud involve tactics such as creating false accounts, using false identities, or manipulating account records. Identifying such fraudulent bank accounts has been an ongoing issue, and numerous detection techniques have been adopted in the process. Within the financial industry, a combination of geographic data, abnormal account activities, and account holder information is used to detect fraudulent accounts. However, the expansion of banking data and an advanced form of bank fraud, known as Synthetic Identity Fraud, has introduced the need for a more capable detection method

Synthetic Identity Fraud involves stealing personally identifiable information of multiple vulnerable individuals, such as young children and elderly adults, and combining them to create a new synthetic identity. New identities are subsequently used to open fraudulent bank accounts and lines of credit. Because these new identities are realistic and target unsuspecting victims, detecting accounts opened with their identities becomes difficult. While the social impacts of Synthetic Identity Fraud are apparent, the victims of synthetic identity theft are not the only ones who suffer. According to a report by the US Federal Reverse, Synthetic Identity Fraud has become the fastest-growing financial crime in the United States, causing financial institutions to lose approximately 6 billion dollars annually from fraudulent bank accounts. This means that regular citizens who trust their money with banks are also affected.

1.2 Prior Work

Both the social and financial implications make the development of reliable and precise detection procedures highly sought after. Researchers have explored a multitude of predictive methods in their pursuit of fraud detection. Methods such as comparative analysis, ensemble learning, and model staking/meta-classifiers have all been used in this task. Additionally, among the variety of metrics used in the evaluations, True Positive Rate and Balanced Accuracy have stood out as the most reliable metrics because they are unaffected by highly imbalanced data, such as real-world fraud detection data.

Although researchers have discovered the best-performing models and the most representative evaluation metrics for fraudulent bank account detection, the issues of reliability, transparency, and explainability remain for practical users such as banks and credit unions. Client privacy and security protocols constrain machine learning processes because they limit the utility of many datasets, leaving very little room to build models with strong explainability and accuracy. Fortunately, a financial crime prevention company, FeedzAI, developed a new type of bank fraud data, which collaborated with researchers from the University of Porto in Portugal. The Bank Account Fraud (BAF) data set applies differential privacy techniques to real-world banking to obfuscate sensitive information while preserving the meaning of attributes, an uncommon yet vital characteristic of fraud data sets. This unique type of fraud data presents an opportunity to develop an accurate and explainable machine learning model. We plan to capitalize on this unique data to build a high-performance model that can detect fraudulent bank accounts with high accuracy, confidence, and transparency.

2 Methods

2.1 Exploratory Analysis

Due to our limited contact with the BAF dataset, a robust exploratory analysis was necessary to become acquainted with the data before modeling. We deployed many techniques to summarize and visually explore our data and found one important insight: The BAF dataset contains less than one percent fraud records; approximately 100,000 fraud records out of 1 million total records. This extremely high class imbalance prompted us to explore if fraud detection is possible through anomaly detection. We assumed that an outlier detection algorithm could easily identify the small proportion of fraudulent bank account records in our data. To test this assumption, we decided to use the Isolation Forest anomaly detection algorithm because of its intuitive approach and ability to handle large data such as ours.

Contrary to our initial assumption, the anomaly detection results show a poor ability to classify fraudulent accounts. The model had a balanced accuracy score of 0.52 and an F-score of 0.14, both of which highlighted the poor classification ability. We examined the results further by visualizing our data. The vast number of features required us to reduce the dimensionality and represent our features using principal components. This enabled us to visualize the components in a conventional scatter plot. We found that there is a lack of defined spatial separation between fraudulent and non fraudulent accounts. In other words, it was impossible to cluster the classes separately in the feature space. These results indicated that fraudulent accounts are surprisingly similar to non fraudulent counterparts and that more complex classification algorithms are required for the task.

2.2 Data Preparation

The extreme class class imbalance we found during EDA was a major concern to our analysis. We were aware that evaluating classifiers is much more difficult with a major class imbalance, as such it was a priority to find an appropriate data balancing method. We experimented with two balancing techniques, over-sampling and under-sampling.

Initially we were concerned that under-sampling would limit the amount of data available for training and testing, so we began with an oversampling technique known as Synthetic Minority Oversampling (SMOTE)3. The procedure works by creating linear combinations of fraudulent records to generate more samples for the minority class (Fraud). While we were successful at creating new synthetic samples, further examinations suggested that the synthesized fraud records were too similar to each other and didn't adequately capture the patterns found in real fraud records. This prompted us to use conventional under-sampling which takes a representative sample of Non-Fraud records to match the number of fraud records in our data. After some initial testing, we decided to move forward with the under-sampled data for the rest of our analysis.

After creating a balanced dataset using under-sampling, we split the data into training and testing sets. We choose a 70/10/20 train/validate/test split to have adequate data for training and testing. Additionally, we replicated the testing set used in previous research papers that explored the BAF dataset, allowing us to compare our classification results to previous works.

2.3 Predictive Modeling

After completing data processing, our goal was to build and compare three types of models that we found worked best during our literature review. They are logistic regression, random forest, and deep neural network. These models increase in complexity, which we believe would enrich our results.

The first model we developed was a Logistic Regression classifier. This is the simplest, and it functions as a baseline for the rest of our models. Our model is trained using the LogisticRegression class from Scikit Learn. It uses an L2 regularization "penalty" to prevent overfitting and the "sage" as our optimization algorithm. We chose Logistic Regression because it is simple to implement and can reveal valuable insights about the features that are most important to our target variable as well as the direction of association between the features and the target.

The second model we built was a Random Forest classifier. It involves building multiple random sized decision trees trained with random feature subsets. This variety allows the model to capture many intricate patterns with in our data. Our model was also trained using the RandomForestClassifier class from Scikit Learn. It uses 150 "tree estimators" and "entropy" for as the optimization criterion in combination with all other default parameters. We chose Random Forest because the feature importance insights that can be easily interpreted and verified.

The final model we developed was a type of Neural Network known as Multi-Layer Perceptron (MLP) from TensorFlow. This model is an ideal fit for tabular data such as the BAF dataset which is composed of numerical attributes and a few categorical attributes that are numerically encoded. After selecting the final model, we conducted some initial testing to explore the ideal model structure for our data. We found an MLP with four total layers, (one input, two hidden, one output), was an ideal structure to balance between predictive ability and model complexity.

As mentioned previously, evaluating fraud classifiers can be challenging due to the major class imbalance. As such, it's vital to asses each of our models using metrics that are not influenced by data imbalance. There are two metrics used in this analysis: Balanced Accuracy and True Positive Rate at a fixed 5% False Positive Rate. The former is the sum of True Positive and True Negative Rates divided by two. While the later is the model's True Positive Rate at the threshold of 0.05 False Positive Rate.

Balanced Accuracy:

$$\text{Balanced Accuracy} = \frac{\text{TPR} + \text{TNR}}{2} \tag{1}$$

True Positive Rate Metric:

$$\text{TPR}_{\text{FPR}=0.05} = \frac{\text{True Positives at FPR} = 0.05}{\text{Total Positives}} \tag{2}$$

2.4 Explainability & Fairness

In order to demystify the black box nature of our best performing model, we need a concept know as explainability. Explainability is a process which reveals how each individual prediction was made, in other words how each feature and its value contributed the predicted probability. There are an variety of explainability tools such as Local Interpretable Model-Agnostic Explanations (LIME). However, in this analysis we focus on SHapley Additive exPlanations (SHAP) for its robust theoretical foundation.
SHAP values come from game theory, and are used to distribute the outcome of a game among its players. For example, if four colleagues join a coding competition and win a $2000 prize, calculating SHAP values would tell us how much of the prize each person has earned based on their contributions. In the context of machine learning, players are replaced with features and the game outcome is replaced with the model's prediction (predicted probability). Doing so allows us to quantify how much each of the features contributed to the prediction of a single record. SHAP values are calculated through an equation known as marginal contribution.

Marginal Contribution Equation:

$$MC(S, i) = v(S \cup \{i\}) - v(S) \tag{3}$$

Marginal Contribution represents the average difference in the model's output (predicted probability) when a feature is included or excluded from every possible subset of features. SHAP values play vital role in the transparency and verifying insights that we examined.

In addition to explainability, a model fairness assessment using Microsoft's machine learning fairness library, FairLearn will strengthen confidence in our model's predictions. Generally, It's desirable that a machine learning model does not make predictions from biases that may exist in sensitive features such as gender or age. It's especially important when models are used in downstream decision making. Fairlearn has a framework to assess and improve model fairness for this exact reason.

Fairlearn allows us to group records based on a sensitive feature such as gender, examine our model's performance for each group (ie. Male/ Female), and finally calculate group-wise comparisons to see if there is a disparity in model performance between the groups. There are three main group-wise metrics used to measure disparity in model performance, Demographic Equality, Equal Opportunity, and Equal odds. Each metric examines an aspect of the model's performance across each group. The metric that best fit's our data is Equal Odds because it examines whether the True Positive and False Positive Rates are consistent across the groups.

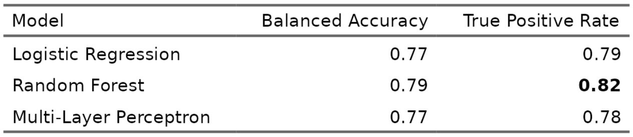
Equal Odds Equation:

$$\max\left(|\text{TPR}(A = a) - \text{TPR}(A = b)|, |\text{FPR}(A = a) - \text{FPR}(A = b)| \right) \tag{4}$$

3 Results

3.1 Classification

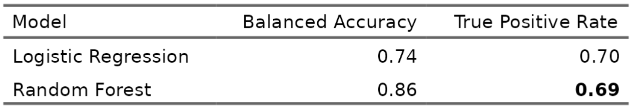
A crucial aspect of our modeling was training with balanced data and testing with unbalanced data that has the original class ratio of 100:1 Non-Fraud to Fraud records. This would allow us to learn without well while ensuring that our model could generalize well to unseen testing data. We began with class ratio 10:1 and gradually moved down to a 1:1 class ratio. We found that as the sampling ratio decreased, the performance on our validation set continued to increase with 1:1 achieving the best performance. These results persuaded us to train all three classifiers with the 1:1 under sampled training data. The figure below shows our results.



Model	Balanced Accuracy	True Positive Rate
Logistic Regression	0.77	0.79
Random Forest	0.79	**0.82**
Multi-Layer Perceptron	0.77	0.78

Figure 2: Fraud Classification Results

While our Random Forest Classifier performed the best in True Positive Rate metric as well as Balanced Accuracy, it's clear that all three of our models have comparable performance. Additionally, because we have replicated testing procedures of previous studies we may compare our results to previous results. The figure below shows previous best performance achieved on the BAF dataset.



Model	Balanced Accuracy	True Positive Rate
Logistic Regression	0.74	0.70
Random Forest	0.86	**0.69**

Figure 3: Best Performance Achieved in Past Studies

Comparing our results to those from previous studies we can see that all three of our models outperformed previous models in terms of True Positive Rate at 5% False Positive Rate, the primary evaluation metric. While our classifiers could not outperform the best Balanced Accuracy, all three models have close Balanced Accuracy scores. We believe that this is acceptable given the improvements to True Positive Rate metric that we were able to achieve.

3.2 Explainability Insights

Calculating SHAP values allows us to quantify how much each of the features contributed to the prediction of a single record. After calculating marginal contribution of each feature and record, the result is a table of SHAP values with the exact dimensions as your original data. This table of SHAP values is crucial in uncovering useful insights through a Feature Contribution plot and Dependence Plots.

The Feature Contribution plot uses what is know as Mean Absolute SHAP value of a feature. It is the sum of absolute SHAP values in the feature's column divided by the total number of records. This represents the total magnitude or impact the feature has on all records and it represents the SHAP equivalent of our Random Forest model's feature importance score. Below are Feature Importance and Feature Contribution plots.

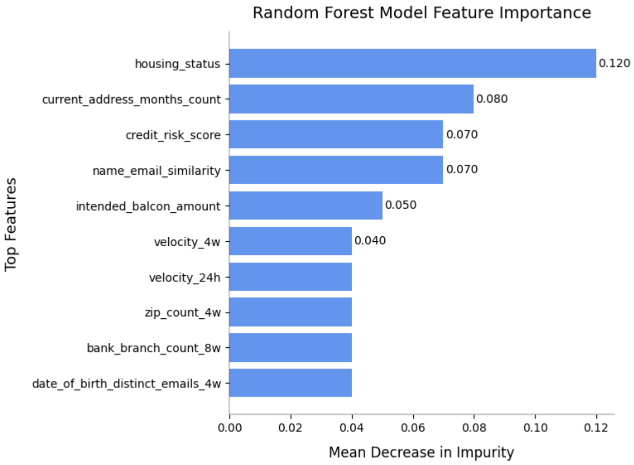


Figure 4: Random Forest Feature Importance Plot

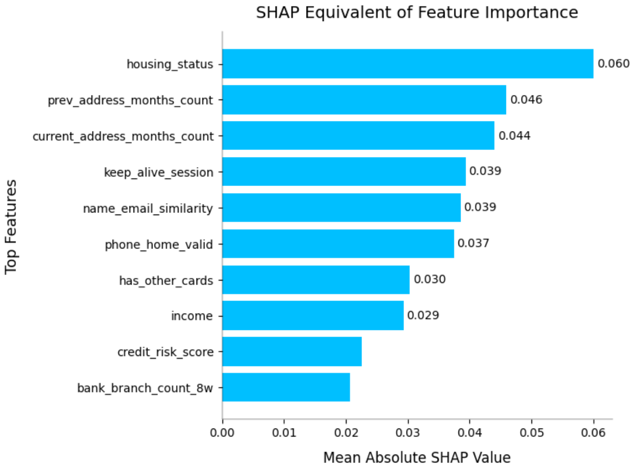


Figure 5: SHAP Feature Contribution plot

Both plots illustrate the top ten most important features respectively. However, the feature importance plot represents the features that are important to our model's decision making during training. The feature contribution plot on the other hand, represents the features that are important to our model when predicting unseen data during testing. The commonalities between the two plots indicate that our model is making predictions as it was trained and intended.

To understand a feature's importance in a model, it is necessary to understand both how changing that feature impacts the model's output, and also the distribution of that feature's values. The dependency plot helps us visualize these relationships. The figure below illustrates the dependence plot for top ten most impactful features.

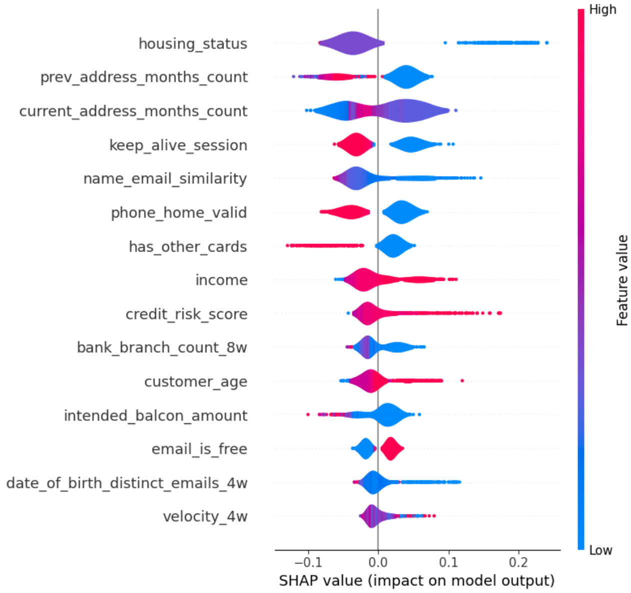


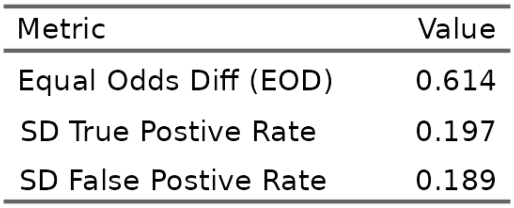
Figure 6: Dependence plot

By examining various features we can see that our model makes assumptions about feature values that are consistent with the way we think of fraud. For example, higher credit risk scores contribute to higher probability of fraud. Similarly, Lower similarity between a customer's name and their email address, the higher the likelihood of predicting fraud. It should be noted that in this context feature values are directly contributing to higher of lower predicted probability. This consistency in assumptions indicates that our model has learned the characteristics of fraudulent accounts.

3.3 Fairness Insights

Unfortunately, we can't develop a model that is unbiased across all the variables that also performs well. This is because our model is classifying through bias and pattern recognition. However, fairness in machine learning means that our model should be unbiased to a few variables such as Gender and Ethnicity. While our data does not contain such sensitive attributes it does a contain Customer Age which we want to mitigate our models bias towards.

To evaluate our models fairness across age groups, we use the Equal Odds Difference metric as it examines whether the True Positive and False Positive Rates are consistent across the age groups. Equal Odds Difference is a value between 0 and 1, the closer your model is to 0, the fairer and unbiased it is deemed. The closer your model is to 1, the more biased it is across your chosen variable. The following is our Random Forest model's Equal Odds Difference.



Metric	Value
Equal Odds Diff (EOD)	0.614
SD True Postive Rate	0.197
SD False Postive Rate	0.189

Figure 7: Equalized Odds Difference & Standard Deviation

The Equal Odds Difference of 0.64 suggests that our model is moderately biased across age groups. However, This does not perfectly encapsulate the fairness of our model. The Standard deviations of True Positive and False Positive Rates indicate that there is much smaller variance than the EOD metric would indicate. We investigated this further and found that the first four age groups contribute the most to the variance in across the groups.

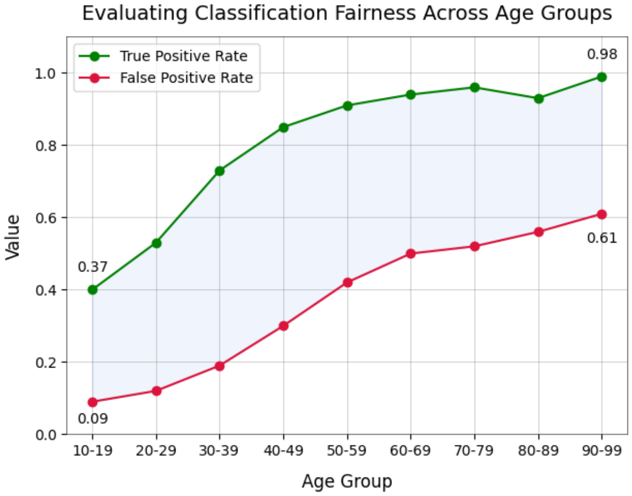


Figure 8: Visualizing Fairness Across Age Groups

This graph shows the True Positive and False Positive rates of the different age groups. If there was no bias, the TPR and FPR lines would be relatively stable and flat, showing similar performance across all age groups. The curvature suggested that there is some bias because of how much the TPR and FPR increased, especially in the first four age groups. This indicates that younger groups have fewer correctly detected fraudulent activity compared to older groups. The plot was created by plotting the EOD of all age groups. Equal Odds examines whether the TPR and FPR are consistent across the groups. This is important to fraud detection since we mainly evaluate TPR at a fixed 5% FPR.

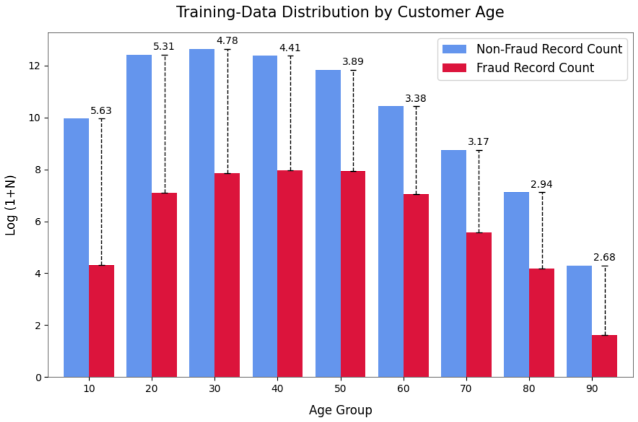


Figure 9: Distribution of Fraud Records

In the graph above we see the training data distribution of the age groups, separated by non-fraud and fraud. Looking more in-depth, we see that the first four age groups do not have many fraudulent records, which is negatively affecting the predictive quality of those age groups. Which means the model has less opportunity to learn meaningful patterns of fraud in younger age groups, potentially reducing its predictive accuracy for these groups. This suggests that the bias is caused by the data's imbalance rather than a flaw in the model.

4 Future Work & Improvements

Fairness and performance

In our project, we addressed the challenge of data imbalance by implementing an improved method of synthetic. The tradeoff between fairness and performance is a critical consideration is machine learning, especially when working with imbalanced data. In our project, as we improved the model, it became more equitable in detecting minority cases. However, this comes to a slight cost to overall performance metrics like accuracy or precision for the majority class. The tradeoff highlights the core tension that optimizing purely for performance can lead to biased outcomes for minority classes. This tradeoff highlights a core tension where optimizing performance can lead to biased outcomes where minority instances are ignored.

Conformal Prediction

Conformal Prediction is used to help us identify which predictions we can trust compared to which ones the model might be unsure of. Conformal Prediction is a statistical framework that generates prediction sets within a specified error rate. Instead of predicting just one class label, conformal prediction will provide a set of predictions (fraud, not fraud, or both) based on a specified confidence level. We completed our code to use the best-performing balanced model (random forest) to perform conformal prediction and measure prediction uncertainty.

We initially chose Inductive Conformal Prediction (ICP) because it is computationally more efficient, however, it does not account for fraud classes. Our data is extremely imbalanced, with very few fraudulent accounts compared to non-fraudulent accounts. Therefore, we think a Mondrian Conformal Prediction approach would be better suited. The Mondrian Conformal Prediction groups data by class, so it would treat fraud and non-fraud cases separately. Thus, we would be comparing "fraud" scores to fraud calibration scores. By combining the best-performing unbalanced model and the calibration dataset, we will be able to generate an adaptive prediction for each prediction in the test data. This directly addresses the idea of making accurate predictions that are supported by robust and reliable prediction sets.

Conclusion

Detecting fraudulent accounts proved to be more challenging than expected, due to its similarity to real genuine accounts, but it is not impossible. However, with proper data preparation and good model selection, we were able to create an ideal Random Forest model that achieved the best performance for our two metrics, TPR at 5% FPR and Balanced Accuracy. When working with fraud detection, it is important to properly balance the dataset that contains meaningful variables, which will boost the performance of the models and help with the explainability and fairness of our predictions.

References

[1] Federal Reserve. (2019). *Payments Fraud Insights: Synthetic identity fraud in the U.S. payment system.* Retrieved from https://fedpaymentsimprovement.org/wp-content/uploads/frs-synthetic-identity-payments-fraud-white-paper-july-2019.pdf

[2] National Credit Union Administration (NCUA). (2018). *Synthetic identities are one of the fastest growing forms of identity theft.* Retrieved from https://ncua.gov/newsroom/ncua-report/2018/synthetic-identities-are-one-fastest-growing-forms-identity-theft

[3] Lepoivre, M. R., Avanzini, C. O., Bignon, G., Legendre, L., & Piwele, A. K. (2016). Credit card fraud detection with unsupervised algorithms. *Journal of Advances in Information Technology, 7*(1), 34–38. https://doi.org/10.12720/jait.7.1.34-38

[4] Kaggle. (2018, March 23). *Credit card fraud detection.* Retrieved from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[5] Jesus, S., Pombal, J., Alves, D., Cruz, A., Saleiro, P., Ribeiro, R. P., Gama, J., & Bizarro, P. (2022). Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. *arXiv preprint.* https://arxiv.org/abs/2211.13358

[6] TransUnion. (2024, February 13). *Banking fraud detection.* Retrieved from https://www.transunion.com/business-needs/fraud-prevention/banking-fraud-detection

[7] Alhashmi, A. A., Alashjaee, A. M., Darem, A. A., Alanazi, A. F., & Effghi, R. (2023). An ensemble-based fraud detection model for financial transaction cyber threat classification and countermeasures. *Engineering Technology & Applied Science Research, 13*(6), 12433–12439. https://doi.org/10.48084/etasr.6401

[8] IEEE Xplore. (2024, November 7). *Evaluation of machine and deep learning methods for fraud detection.* Retrieved from https://ieeexplore.ieee.org/abstract/document/10757257

[9] IEEE Xplore. (2023, January 27). *Fraud detection in banking transactions using machine learning.* Retrieved from https://ieeexplore.ieee.org/document/10091067

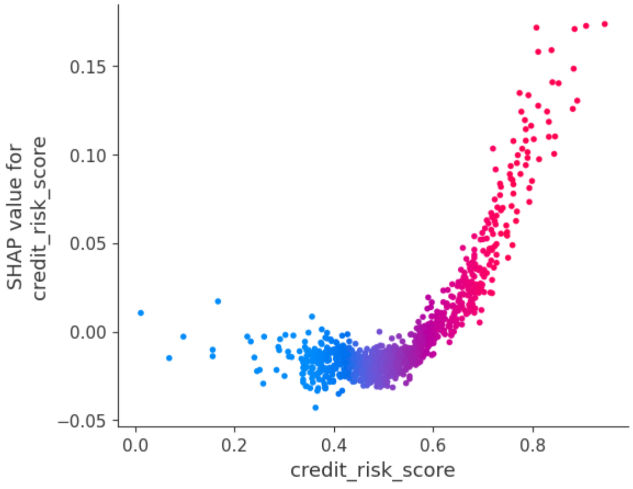


Figure 12: Credit Risk Score Dependence Plot

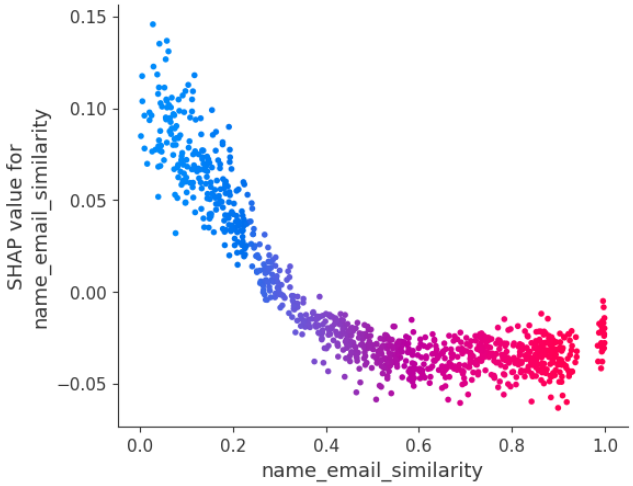


Figure 13: Name Email Similarity Score Dependence Plot

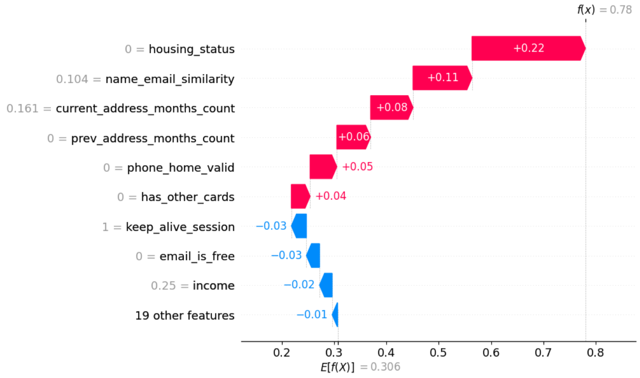


Figure 14: True Positive Prediction Waterfall Plot

[10] Angelopoulos, A. N., & Bates, S. (2017). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint.* https://arxiv.org/abs/2107.07511

[11] Lundberg, S. M. (2017). A unified approach to interpreting model predictions. Proceedings of NeurIPS. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[12] Pombal, J., Saleiro, P., Figueiredo, M. A. T., & Bizarro, P. (2023). Fairness-aware data valuation for supervised learning. *Feedzai & Instituto Superior Técnico, Universidade de Lisboa.* Retrieved from https://arxiv.org/pdf/2303.16963

5 Additional Figures

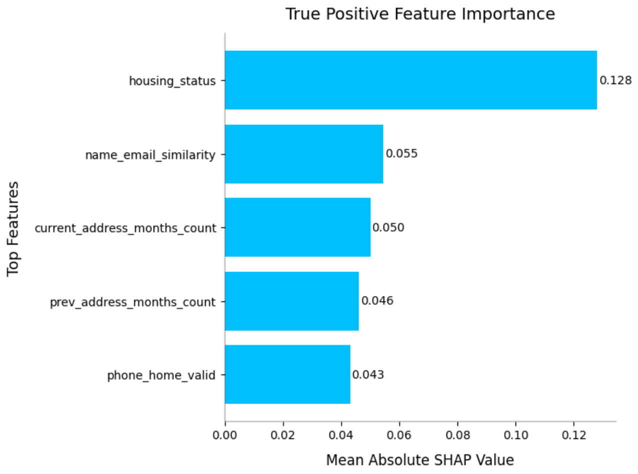


Figure 10: True Positive Feature Contribution

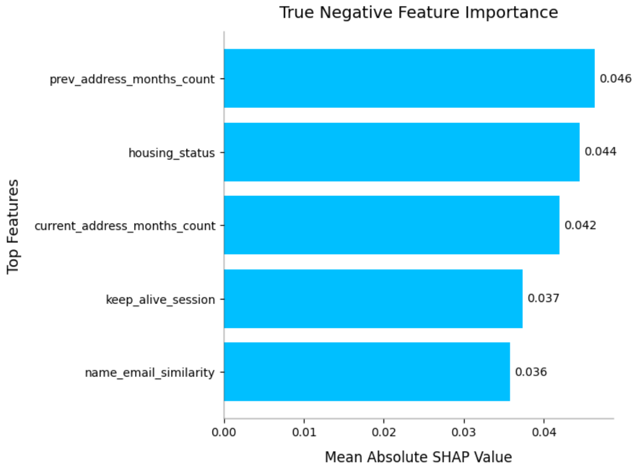


Figure 11: True Negative Feature Contribution