Jose Ayala

Cuneyt Akcora

CAP 5619 - Artificial Intelligence for FinTech

April 19, 2025

Fraud Detection with Data Imbalance

Datasets and Motivation

I analyzed the Credit Card Fraud Detection dataset and the Simulated Credit Card Transactions dataset. My motivation for choosing these datasets was to understand fraud detection in scenarios with varying degrees of class imbalance and to explore model explainability. Both datasets relate to card-not-present fraud.

Model Architecture and Training Setup

For both datasets, we used a simple MLP neural network with an input layer, two hidden layers (64 and 32 neurons, ReLU), and a sigmoid output layer. The Credit Card dataset was split into stratified train/validation (80/20). Models were trained using Adam, binary cross-entropy, batch size 32, and 10 epochs, monitoring AUC, accuracy, precision, and recall. For the Simulated data, we preprocessed categorical features (one-hot encoding), scaled numerical features, and trained on original, oversampled, and undersampled data with similar settings.

Performance Metrics Before and After Balancing

Credit Card Data:

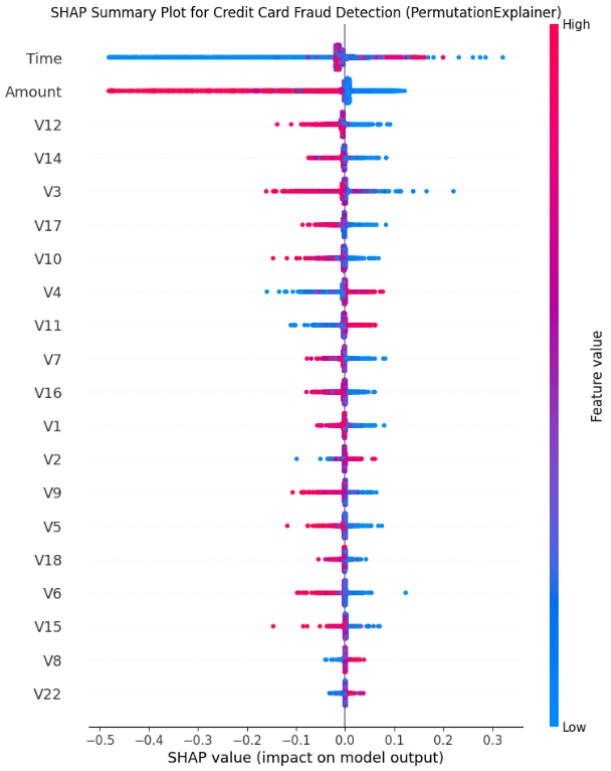
- **Imbalanced**: Poor fraud detection (AUC ~0.5, precision/recall ~0). High accuracy was misleading.

- **Oversampled**: Significant improvement in AUC (~0.95) and recall (~0.89) on the original validation set, but lower precision (~0.04).
- **Undersampled**: Showed modest improvement in AUC (~0.56) compared to the imbalanced data, with a precision of ~0.50 and a recall of ~0.13 on the validation set.

Class imbalance severely hindered the initial model's ability to detect fraud. Balancing techniques significantly improved fraud detection capability, with oversampling leading to higher AUC and recall at the cost of precision, while undersampling provided a more balanced, albeit still modest, improvement in these metrics.

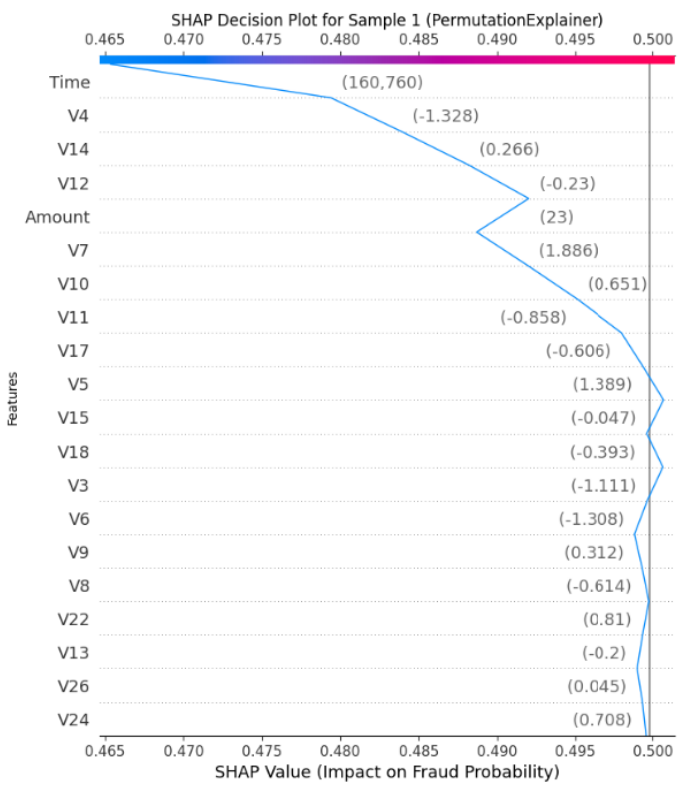
SHAP Analysis and Visualizations

The SHAP summary plot for the oversampled Credit Card model revealed the most influential features to be Time and Amount.



Individual SHAP decision plots showed how specific feature values contributed to the model's prediction for a given transaction, illustrating the push towards or away from a fraudulent classification.

Explanation for prediction on test sample 1:




```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 49ms/step
  Predicted Probability of Fraud: 0.4654
  Predicted Class: Not Fraudulent
  Actual Class: Not Fraudulent

The Top 3 Most Influential Features on this sample include:

  1. Feature 'Time':
    Impact: 0.0141 (had a negative influence)
    Effect: This feature decreased the predicted probability of fraud.
    Note: This was the most influential feature for this prediction.
--------------------------------------------------------------------------------
  2. Feature 'V4':
    Impact: 0.0045 (had a negative influence)
    Effect: This feature decreased the predicted probability of fraud.
--------------------------------------------------------------------------------
  3. Feature 'V14':
    Impact: 0.0043 (had a negative influence)
    Effect: This feature decreased the predicted probability of fraud.
--------------------------------------------------------------------------------
```

Reflection on Class Imbalance and Insights

Our analysis of both the Credit Card Fraud Detection and Simulated Credit Card Transactions datasets highlighted the significant impact of class imbalance on model performance. In both cases, models trained on the original imbalanced data struggled to effectively identify fraudulent transactions, demonstrating a strong bias towards the majority class. Addressing this imbalance through oversampling and undersampling proved crucial for improving fraud detection capabilities, leading to better AUC and recall, with trade-offs in precision. This highlights the need to employ class balancing techniques when dealing with rare event prediction.

Explainability tools like SHAP offered valuable insights into the models' decision-making processes for the Credit Card Fraud Detection data. SHAP revealed the importance of the anonymized features in driving the model's predictions. Analyzing individual predictions through SHAP decision plots illustrated how specific values of these features influenced the likelihood of a fraudulent classification. These explainability techniques provide a degree of transparency, helping to understand the models' reasoning and identify key drivers of fraud predictions, ultimately contributing to building more reliable and interpretable fraud detection systems for credit card fraud.