Fraud Data Scientist Case

Andressa Ribeiro
e-mail: andressaribeiro@usp.br

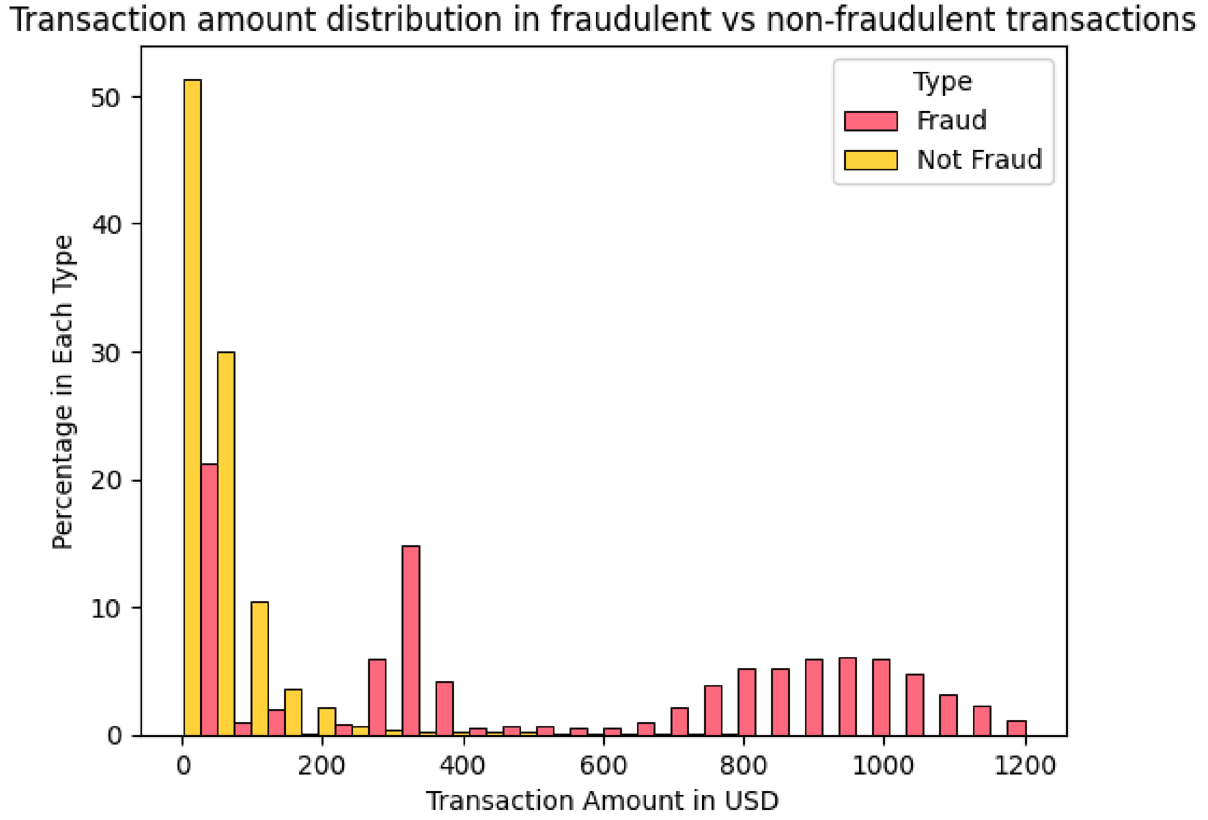
PROBLEM STATEMENT

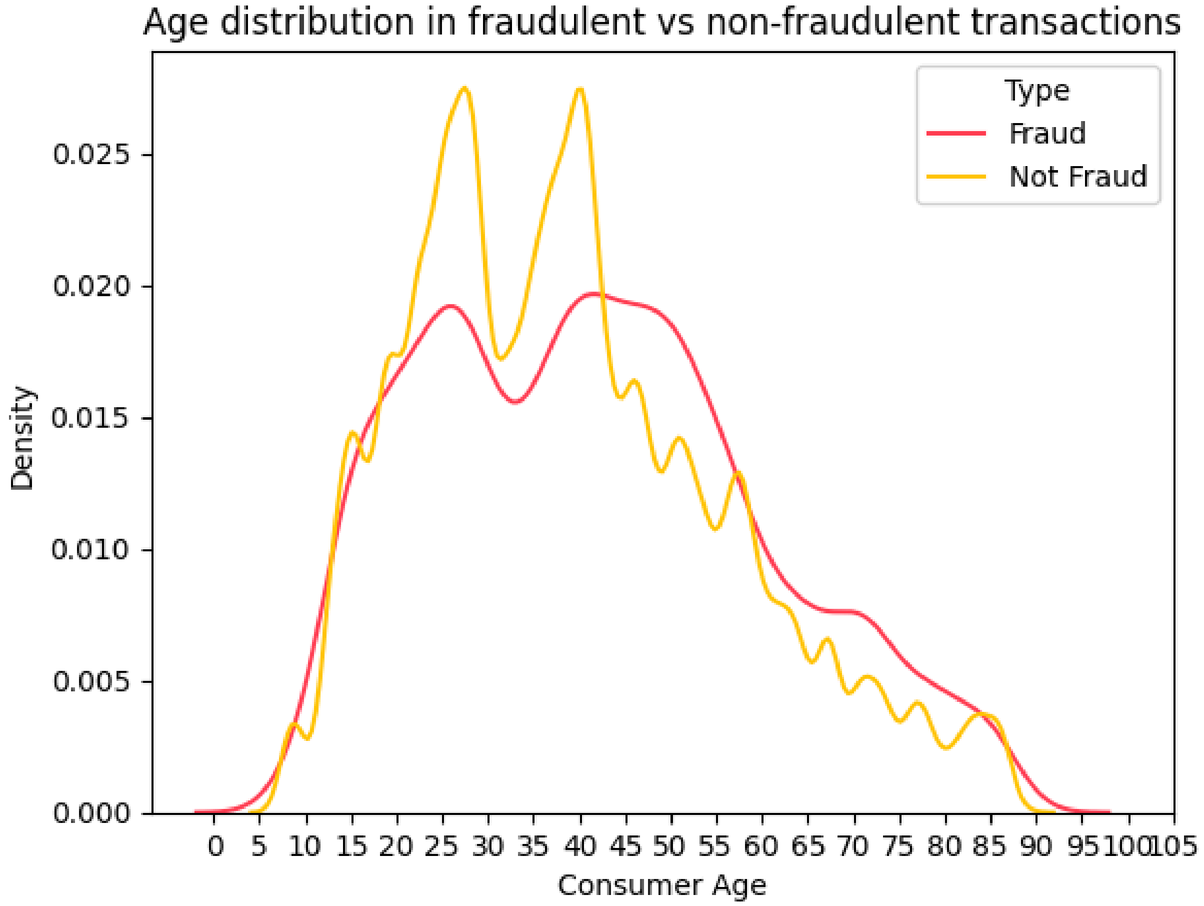
CREDIT CARD TRANSACTIONS

- Source: Kaggle - Fraud Detection Dataset.
- Objective: Build a robust ML model in Python to improve the fraud detection system.
- Description: Credit card transaction dataset labeled with **is_fraud**.
  - Structure:
    - fraudTrain.csv: For model training and refinement.
    - fraudTest.csv: Exclusively for model performance evaluation.

Volume of transactions: 1,852,394

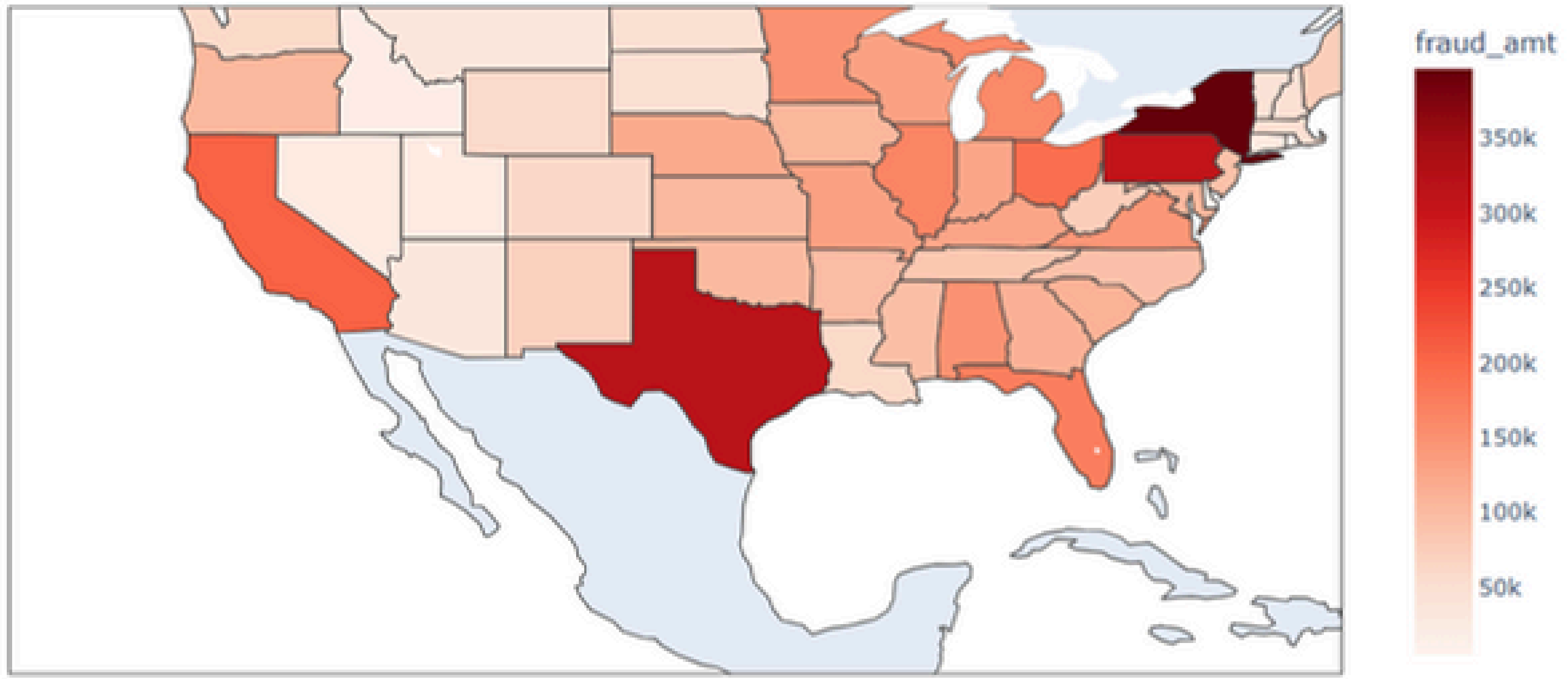
EXPLORATORY DATA ANALYSIS

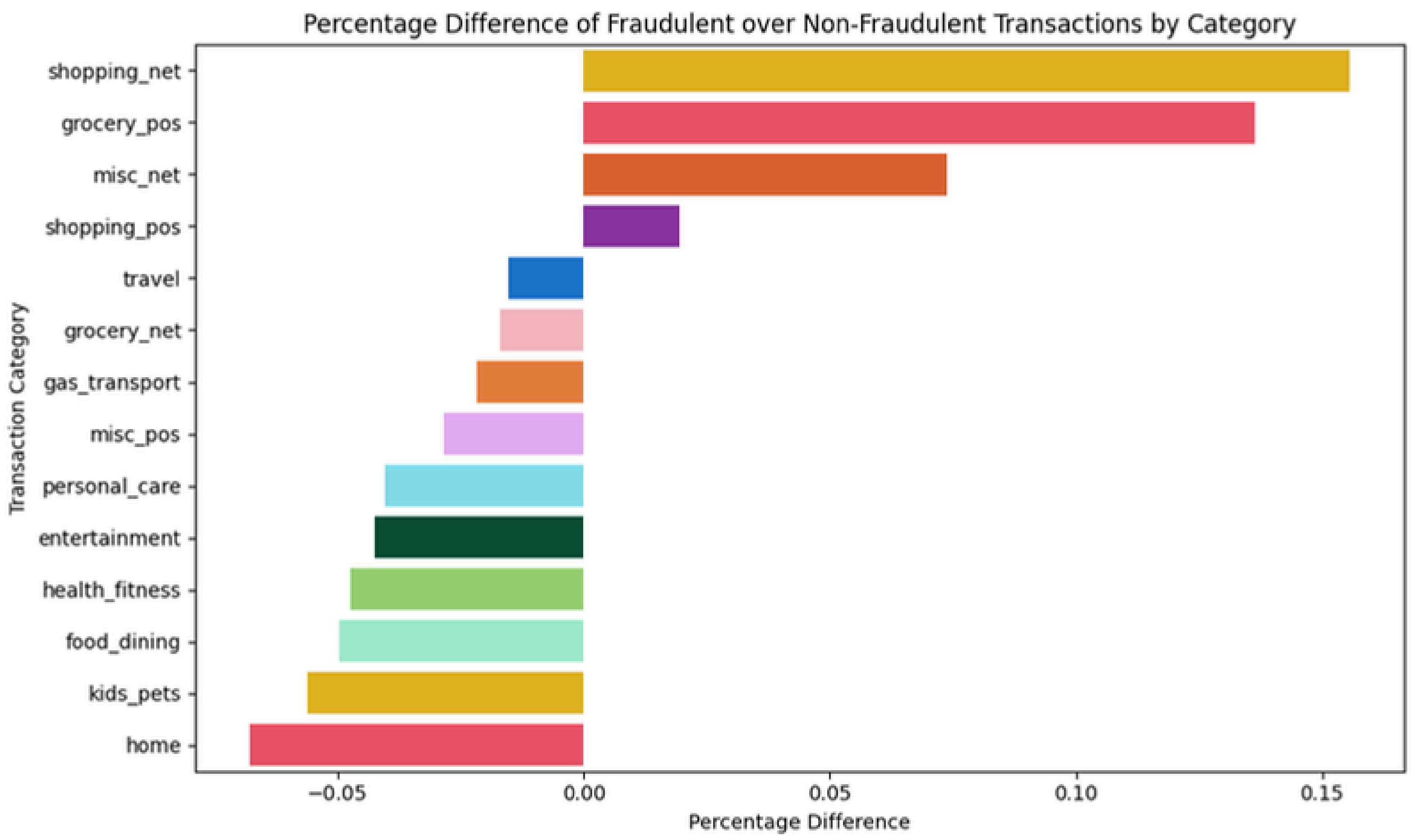




- Higher transaction values show a stronger relationship with fraudulent activity.
- Fraud concentrated in 30-50 age range, possibly due to higher value/volume transactions.
- Legitimate transactions peak among 20-35 year olds.
- From ~40 years onwards, fraudulent transaction density surpasses legitimate ones.

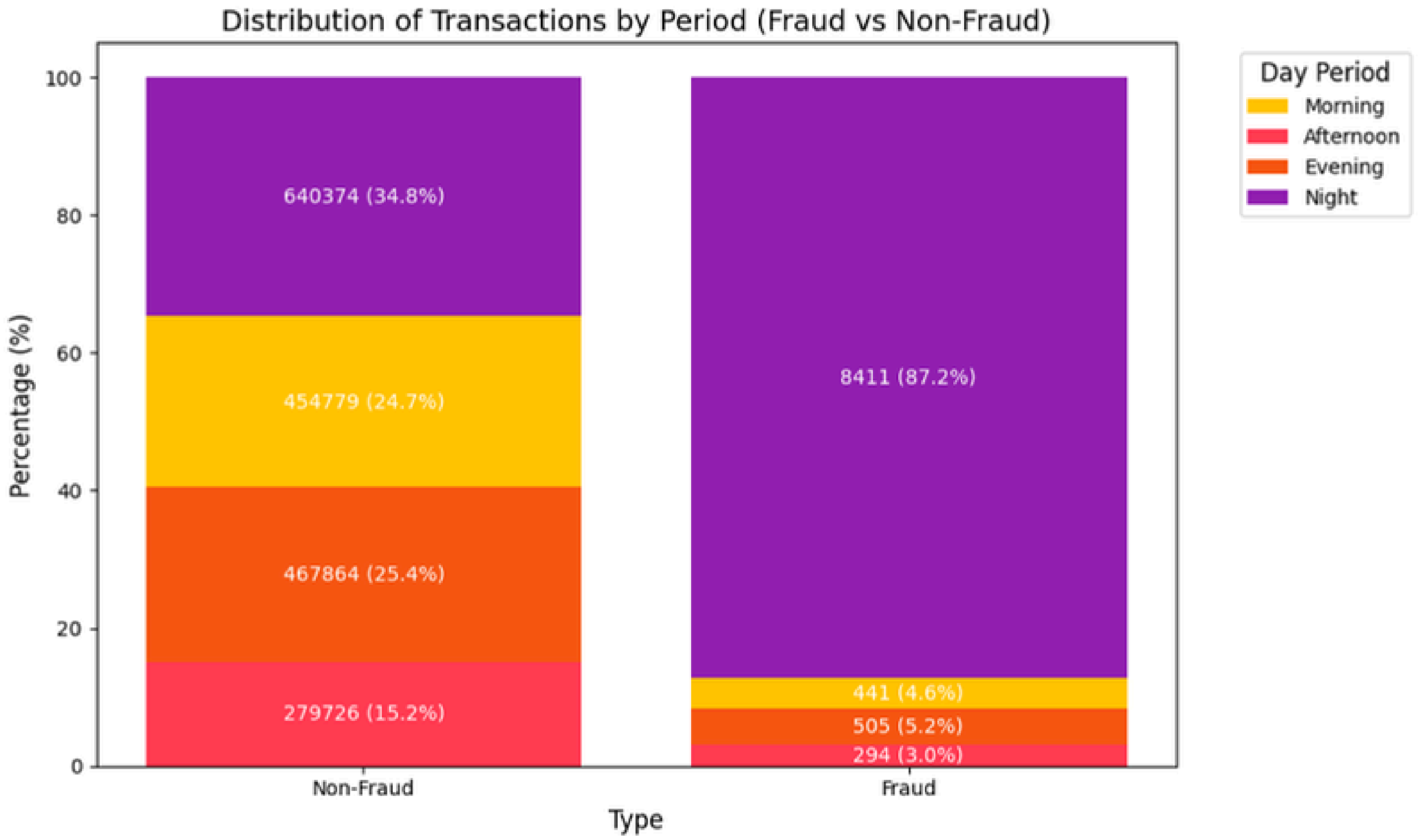
EXPLORATORY DATA ANALYSIS

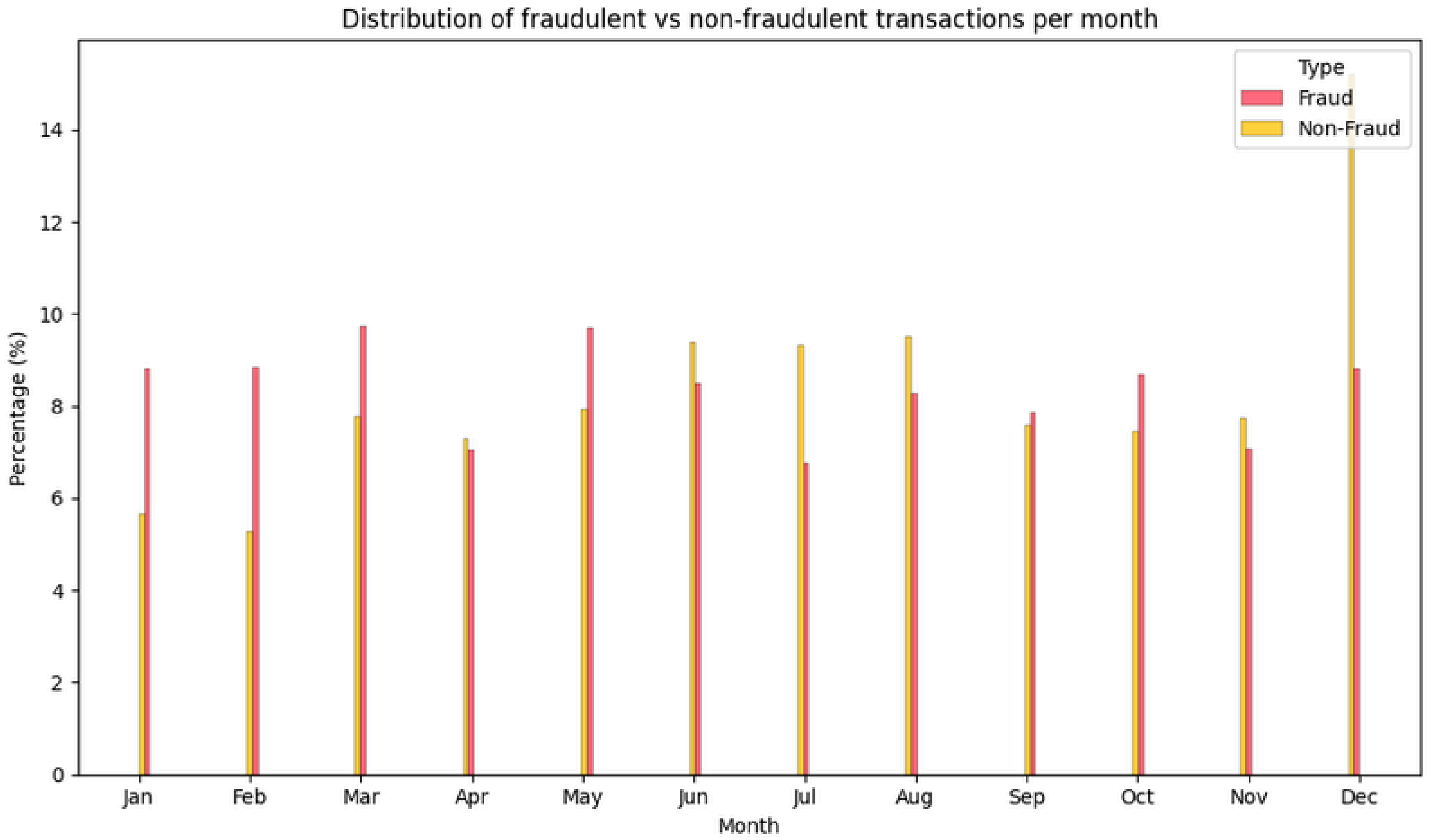




- When considering the total fraudulent amount, states like New York, Pennsylvania, and Texas stand out.
- Overall, the Midwest and Northeast regions show a higher concentration of fraud compared to other areas.
- The most frequently defrauded categories are shopping, miscellaneous, and groceries.

EXPLORATORY DATA ANALYSIS





- Fraudulent transactions tend to occur more frequently during early morning hours (overnight).
- There's also a pattern of increased fraud at the beginning of the year and specifically between Wednesday and Friday.
- Addressed high-cardinality features (merchant, city, job) by prioritizing grouped features or advanced encoding to prevent overfitting.

MODELING

- Tackled severe class imbalance with SMOTE + Undersampling, resulting in a balanced training set of 156,540 samples (41% fraud).
- Conducted rigorous hyperparameter tuning across multiple algorithms (e.g., LightGBM, XGBoost) using 8-fold Cross-Validation, optimizing for AUC.
- Ensured model generalization and stability through Nested Cross-Validation (3 outer / 5 inner folds) and a final 10-fold CV on the best model's parameters.

Algorithm	**AUC**	**Recall**
Logistic Regression	0.960	0.825
Decision Tree	0.952	0.771
Random Forest	0.963	0.790
XGBoost	1.000	0.991
LightGBM	1.000	0.991

PREDICTING

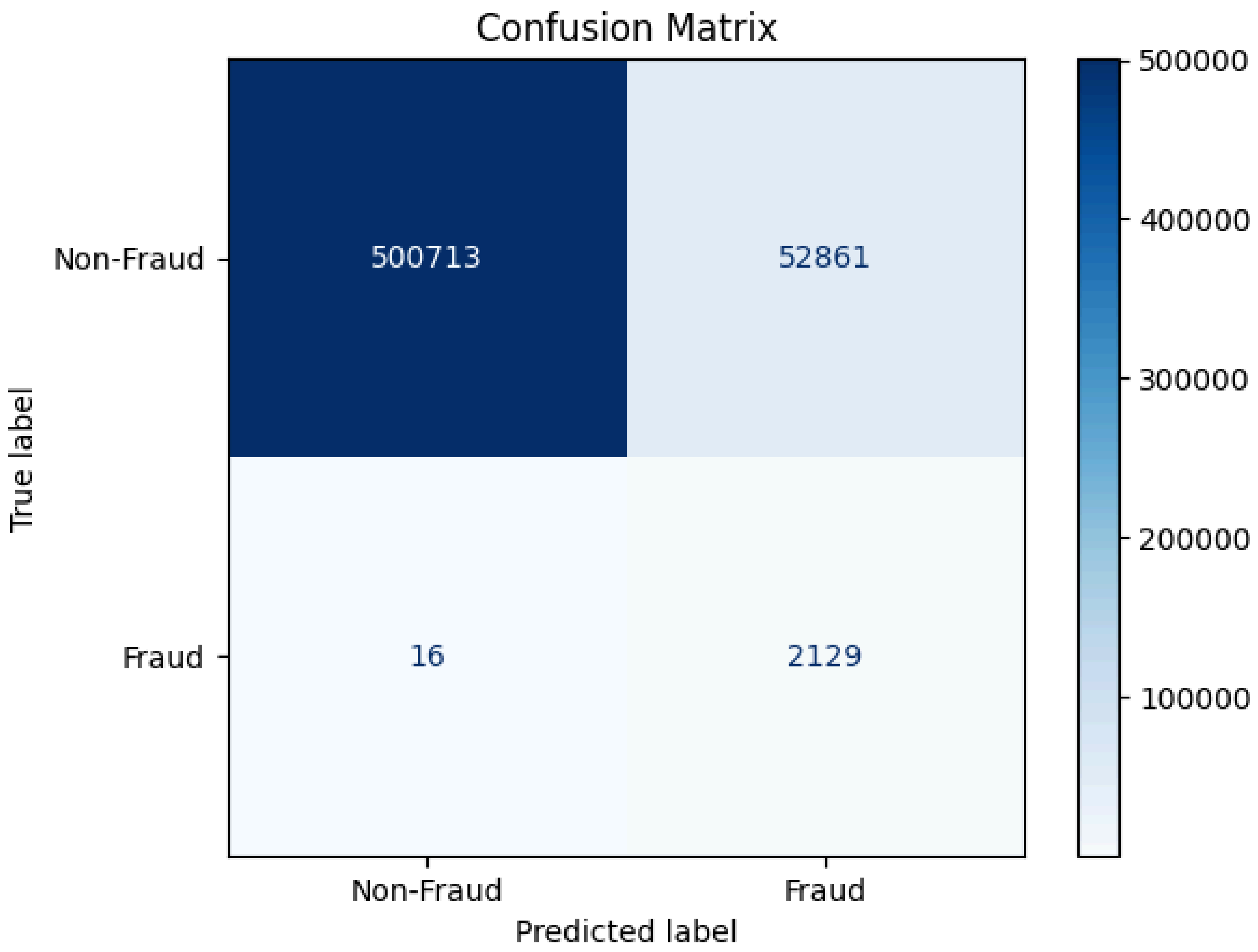
Test metrics:

F1-Score: 94.65%

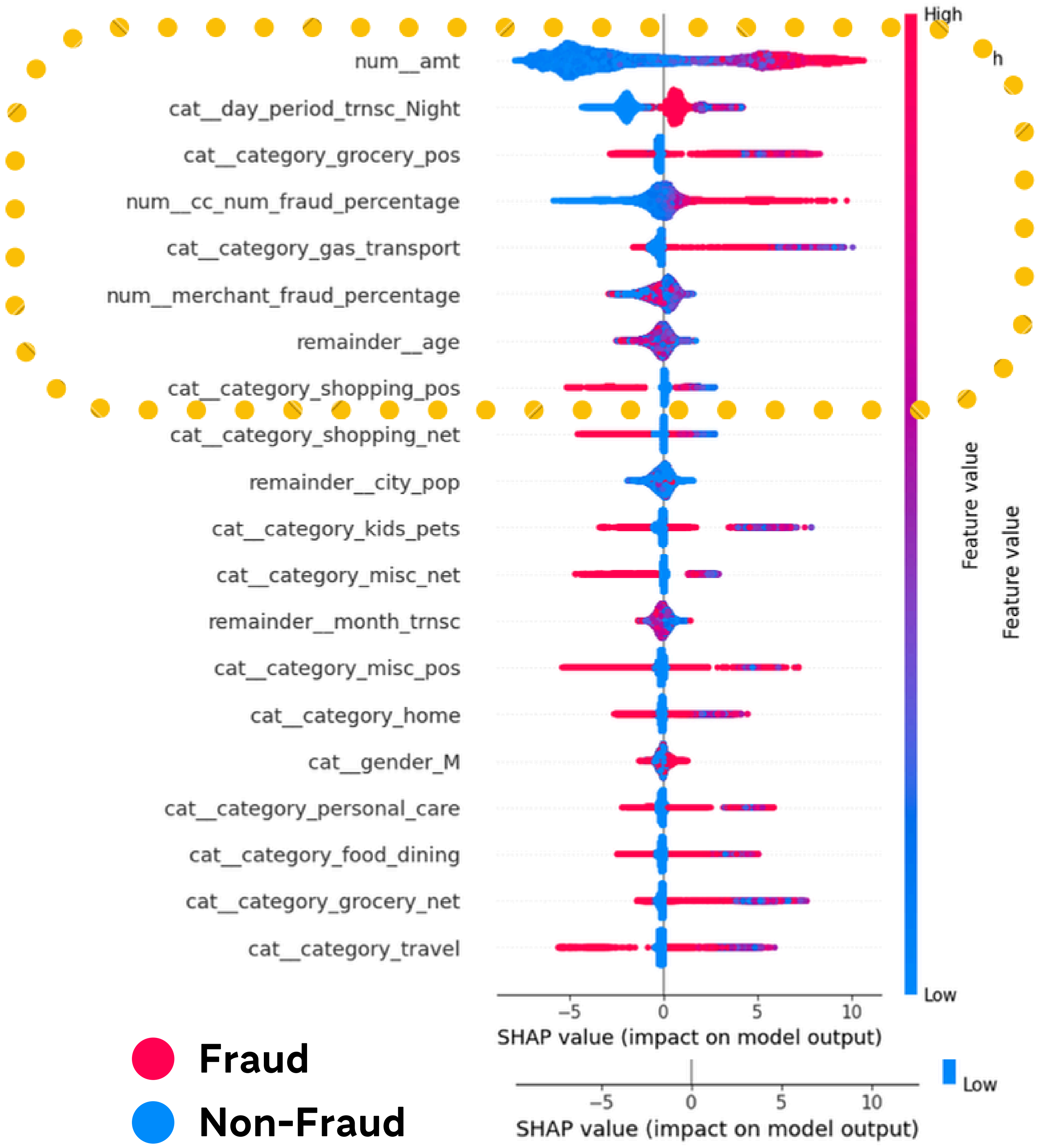
True positive (TP): 2,146
False positive (FP): 16
False negative (FN): 52,861

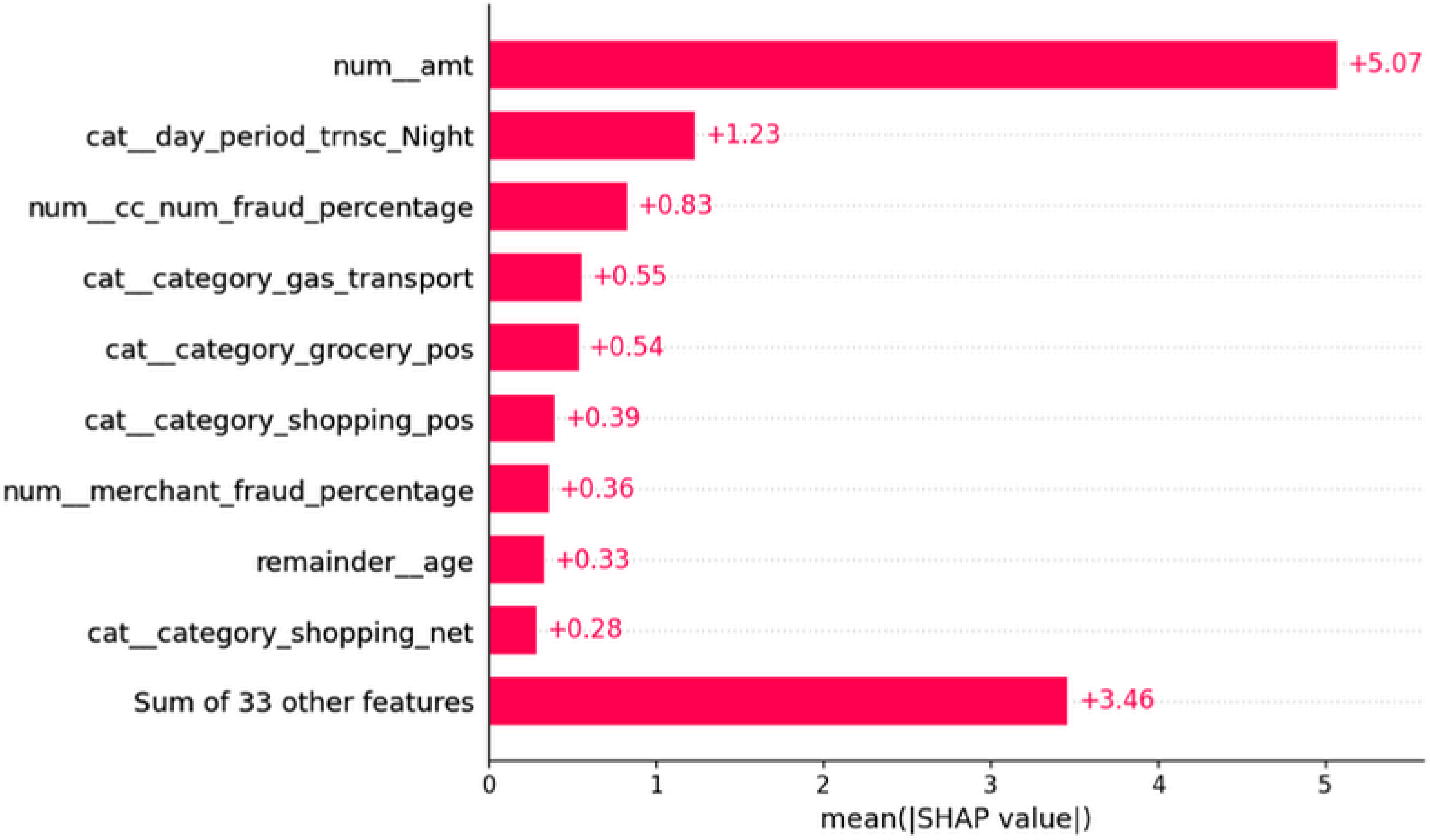
Recall: TP/(TP+FP)

After model: 99.25%
No baseline available



PREDICTING





Fraud

Non-Fraud

FINANCIAL RESULTS

NET FINANCIAL IMPACT

WITHOUT MODEL	
Potential Fraud Loss (R$)	-R$ 793,327.28
Lost Revenue from False Positives (R$)	R$ 0
Avoided Fraud Loss (R$)	R$ 0
Net Financial Impact (R$)	**-R$ 793,327.28**

WITH MODEL	
Potential Fraud Loss (R$)	-R$ 475.26
Lost Revenue from False Positives (R$)	-R$ 306,928.35
Avoided Fraud Loss (R$)	+R$ 792,852.02
Net Financial Impact (R$)	**+R$485.448,41**

+R$ 1,278 K

THANK YOU!

Andressa Ribeiro
e-mail: andressaribeiro@usp.br