Detecting Fraudulent Transactions Using Machine Learning

FATIMA CAMELA RAMOS

JAN 31, 2026

Suspected Digital Fraud Rate (~13%) in the Philippines remains high for five consecutive years

7 out of 10 Filipinos reported being targeted by email, online, phone call or text messaging fraud recently. Among Filipinos who said they lost money due to fraud recently, the reported average loss is around Php45,000. This amount is equivalent to a little over 2 months worth of salary for a minimum wage earner in NCR. (TransUnion, 2025)

While fraudulent transactions account for a small fraction of total transaction volume. even a single successful fraud incident can result in disproportionate financial, emotional, and operational costs

Fraud incidents also impose emotional stress on customers and increasing regulatory and compliance costs on financial institutions

The challenge is to be able to detect high-risk behaviors of fraudulent transactions before damage occurs

Problem Statement

Financial fraud detection in digital transactions remains challenging due to the **rarity of fraudulent events** and the **continually evolving behavioral patterns** employed by threat actors.

In highly imbalanced datasets, traditional linear and rule-based approaches often struggle to reliably distinguish fraudulent behavior from legitimate transactions, resulting in missed fraud cases and excessive false alerts.

Objective

This study aims to **evaluate machine learning models and identify relevant behavioral and contextual signals** that effectively prioritize fraudulent transactions under real-world imbalance constraints.

Methodology

Dataset — Primary

Historical transaction data is used to identify patterns associated with fraudulent behavior

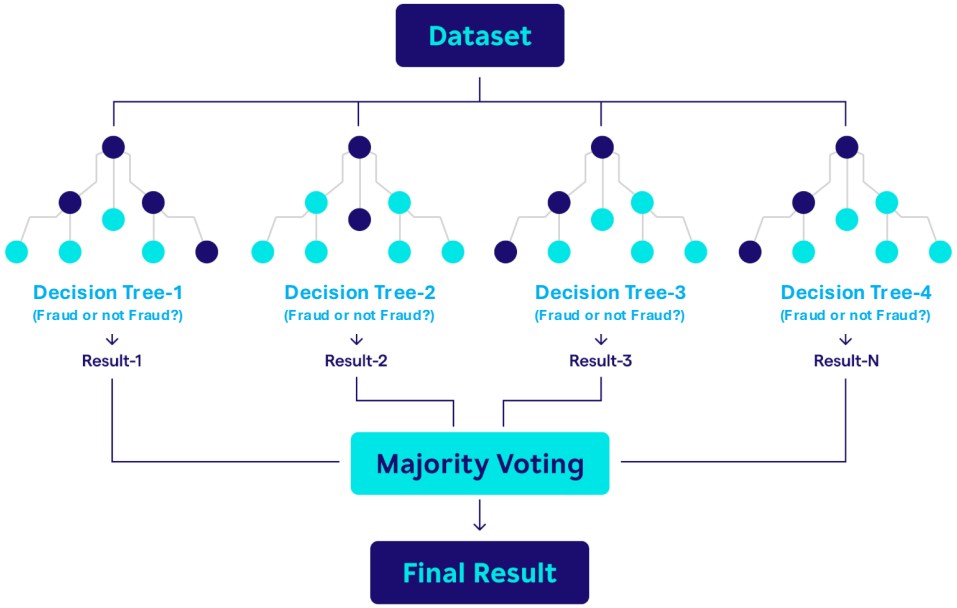
Feature Engineering

The variables were modified to enable the machine to identify patterns more efficiently

Train-Test Split

Transformed data was divided into two sets: one set will be used to train the model, while another set will be used to test model performance vis-à-vis success metrics

Random Forest



Model Evaluation — Confusion Matrix, PR-AUC, SHAP

Model performance is reviewed vis-à-vis target metrics, and parameters are adjusted to improve performance when necessary.

External Validation — External

The model was tested using another dataset with similar features and use case to check if it can still perform well

Datasets Used

Primary Dataset

The primary dataset used to build the model is a synthetically generated data with 5 million transactions mimicking real-world scenarios. It was selected primarily for its scale and feature richness. ~3-4% of transactions are fraudulent. This dataset was used for feature engineering, model development, and internal evaluation:
https://www.kaggle.com/datasets/aryan208/financial-transactions-dataset-for-fraud-detection

External Validation Dataset

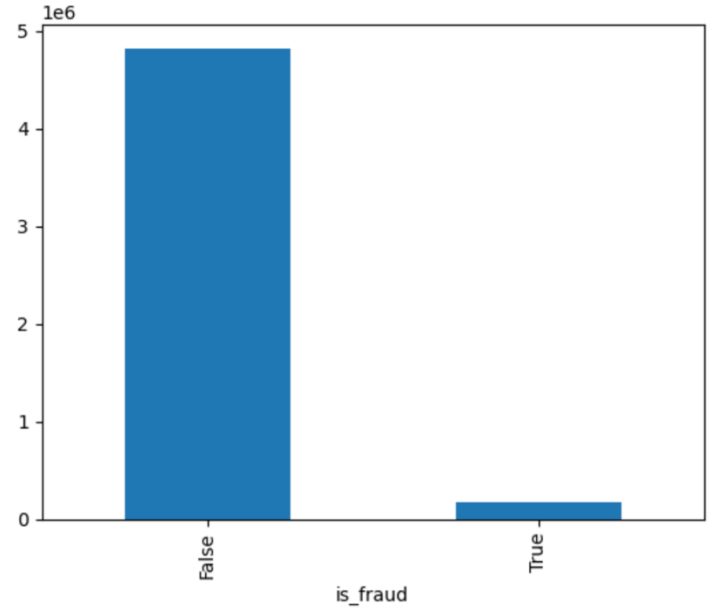
To test how the model performs in real world datasets, external validation was conducted.
This dataset presents actual transactions of European cardholders that occurred in two days in 2013. 0.17% of the transactions are fraudulent. Only overlapping features (e.g., transaction timing, amount, fraud label) were used to ensure comparability during external validation.
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Exploratory Data Analysis

EDA reveals the structure, patterns, and constraints of the data, guiding feature engineering, model choice, and evaluation strategy before formal modeling begins.

Column	Dtype
transaction id	object
timestamp	object
sender account	object
receiver account	object
amount	float64
transaction_type	object
merchant_category	object
location	object
device used	object
is fraud	bool
f raud_type	object
time since last transaction	float64
spending_deviation_score	float64
velocity_score	int64
geo_anomaly_score	float64
payment_channel	object
ip_address	object

The dataset has a total of 18 columns: 1 Boolean, 5 numerical features, and 12 categorical features



The dataset is heavily imbalanced, with 179,553 out of 5M fraud transactions (3.6%)

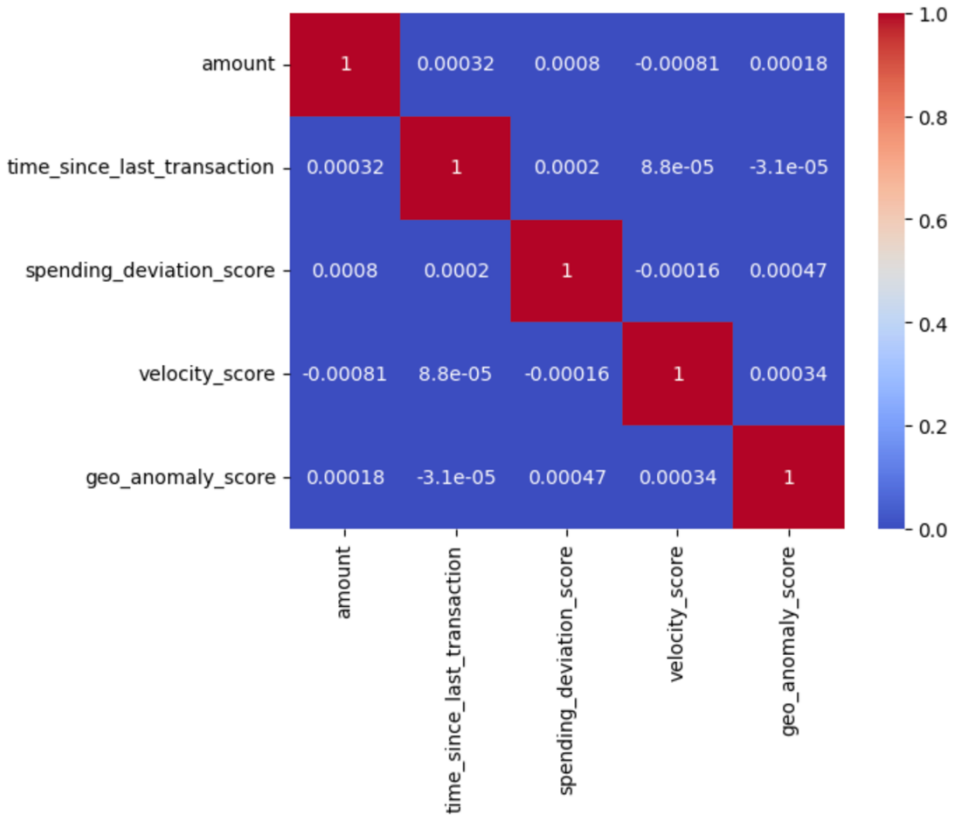
fraud_type	4820447
time_since_last_transaction	896513

Two features contain missing values. The number of fraud_type with missing values is equivalent to non-fraud transactions, therefore, these rows were kept in the dataset and were transformed during pre-processing.

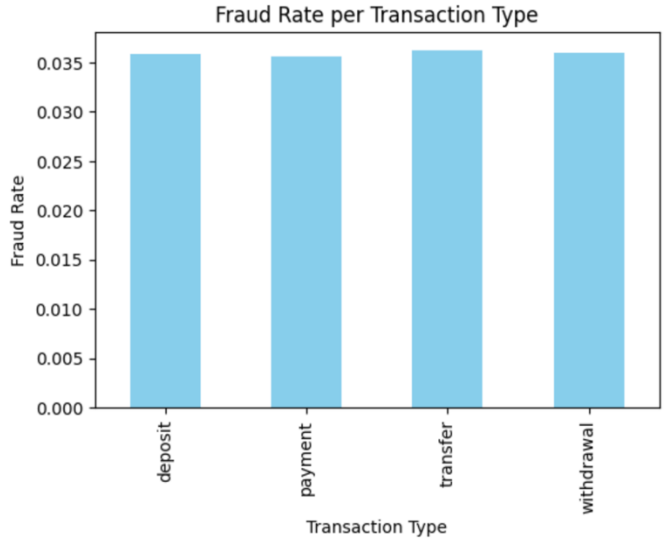
To handle the missing values in time_since_last_transaction, the median of the field was used.

Exploratory Data Analysis

Correlations between variables are very weak, confirming that a linear model will not be able to effectively identify high-risk behaviors




```
transaction_type
deposit       0.035812
payment       0.035640
transfer      0.036253
withdrawal    0.035938
Name: is_fraud, dtype: float64
```



Fraudulent transactions are **evenly distributed across transaction types**, with no single category dominating fraud occurrence.

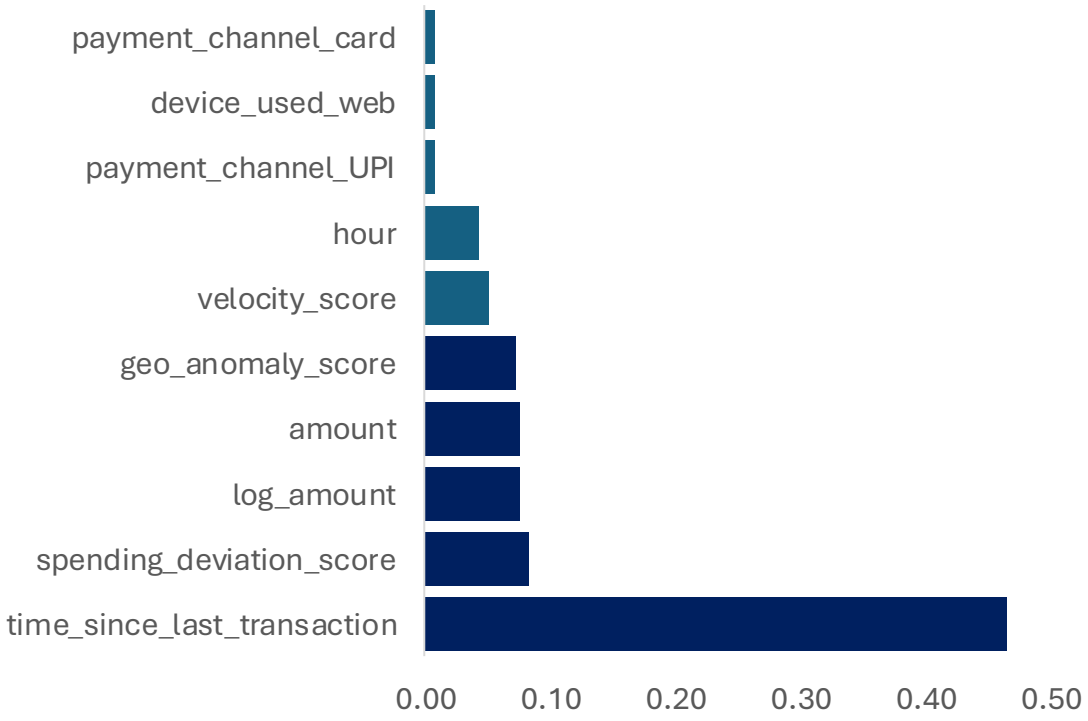
Exploratory Data Analysis indicates the absence of strong multicollinearity across variables, suggesting that feature relationships are not dominated by linear dependencies and supporting the use of **non-linear modeling approaches** (e.g. tree-based models). EDA also reveals that several features exhibit near-uniform distributions, indicating limited standalone discriminatory power. Given that the dataset is synthetically generated, these findings highlight the need to **evaluate the model on real-world transaction data** to adequately stress-test performance and assess generalizability.

Feature Engineering

This step covers transforming raw transaction data into behavioral and contextual features that improve fraud risk discrimination under class imbalance.

Addressed outliers that may impact model training → Created Time Bins to capture potential time of day risk patterns → Hot Encode Categorical Features → Scale features using StandardScaler → Rank features according to importance → Select top 10 features to be used for model training

Which features are important to determine if a transaction is likely to be fraudulent?



From explanatory data analysis of the dataset, the **time between transactions is the strongest indicator** of fraudulent behavior.

The shorter the gap between transactions, the more likely it is to be fraudulent.
Other important features are:

- Spending that seems to deviate from normal spending patterns
- Location where the transaction is done
- Transaction amount

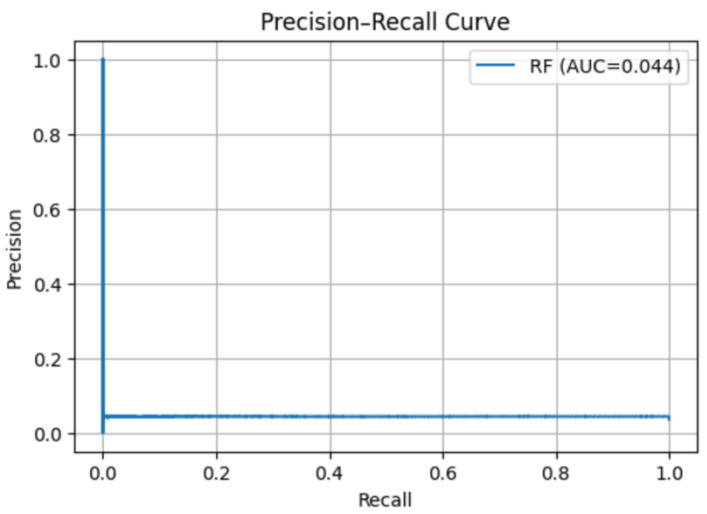
Models Reviewed

Classification models (e.g. logistic regression, random forest, XGBoost) were experimented on to find the most appropriate model.

PR-AUC was selected as the primary metric as it directly evaluates the trade-off between identifying fraud (recall) and minimizing false alarms (precision), which aligns with real-world fraud investigation costs.

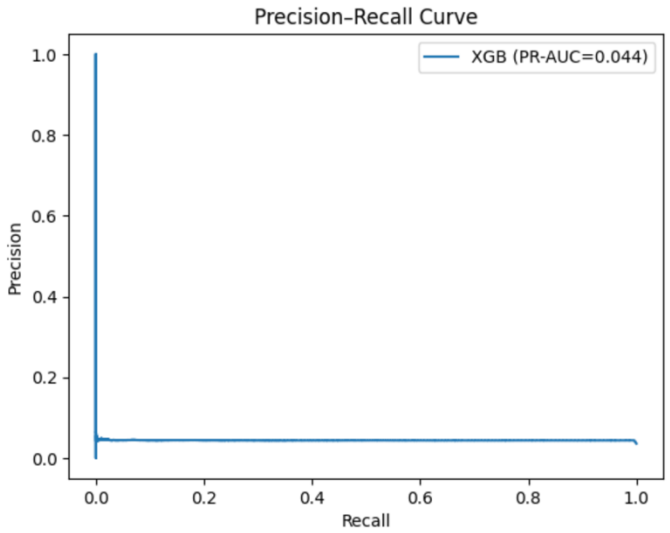
Random Forest was selected due to comparable PR-AUC performance, greater stability under noisy synthetic data, lower overfitting risk, and stronger interpretability for audit and governance use cases.

Random Forest



	precision	recall	f1-score	support
False	0.99	0.21	0.35	964089
True	0.04	0.97	0.08	35911
accuracy			0.24	1000000
macro avg	0.52	0.59	0.22	1000000
weighted avg	0.96	0.24	0.34	1000000

XGBoost Classifier



	precision	recall	f1-score	support
False	1.00	0.20	0.33	964089
True	0.04	0.98	0.08	35911
accuracy			0.23	1000000
macro avg	0.52	0.59	0.21	1000000
weighted avg	0.96	0.23	0.32	1000000

Scope, Limitations, and Ethical Considerations

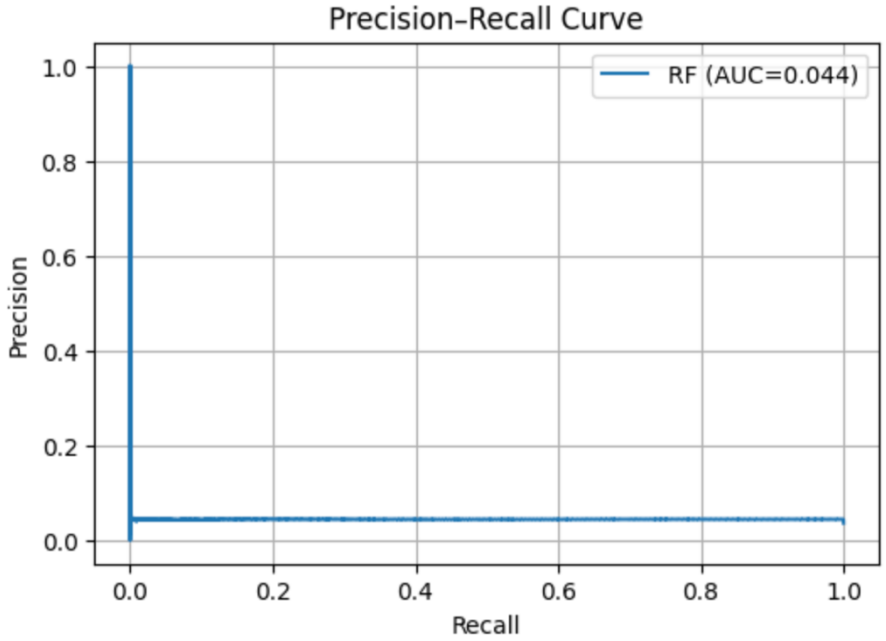
- The historical dataset used to train the model contains approximately **5 million transaction records**, providing sufficient scale to identify behavioral patterns. While the dataset is **synthetically generated**, it was designed to closely mimic real-world transaction behavior.

- The dataset is **highly imbalanced**, with **fraudulent transactions representing only ~3–4%** of total activity. This reflects real-world conditions and informed the use of **classification models optimized for recall and its ability to prioritize real fraud cases over low-risk transactions in a highly imbalanced environment**, rather than accuracy alone.

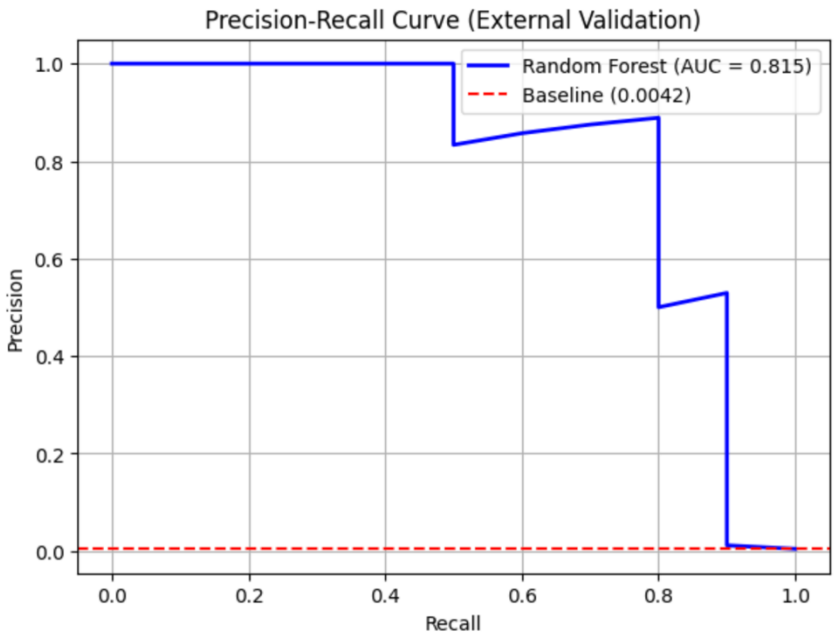
- The dataset does not contain direct personally identifiable information (PII). If future implementations include demographic or sensitive attributes, additional bias monitoring and fairness checks would be required. Human oversight will also be practiced to mitigate bias when selecting features.

- Model behavior is explained using interpretable methods (e.g., SHAP) to support transparency, auditability, and informed human review.

Results and Insights



Model prioritization marginally exceeds baseline risk ranking on synthetically generated dataset (0.04 vs 0.036)

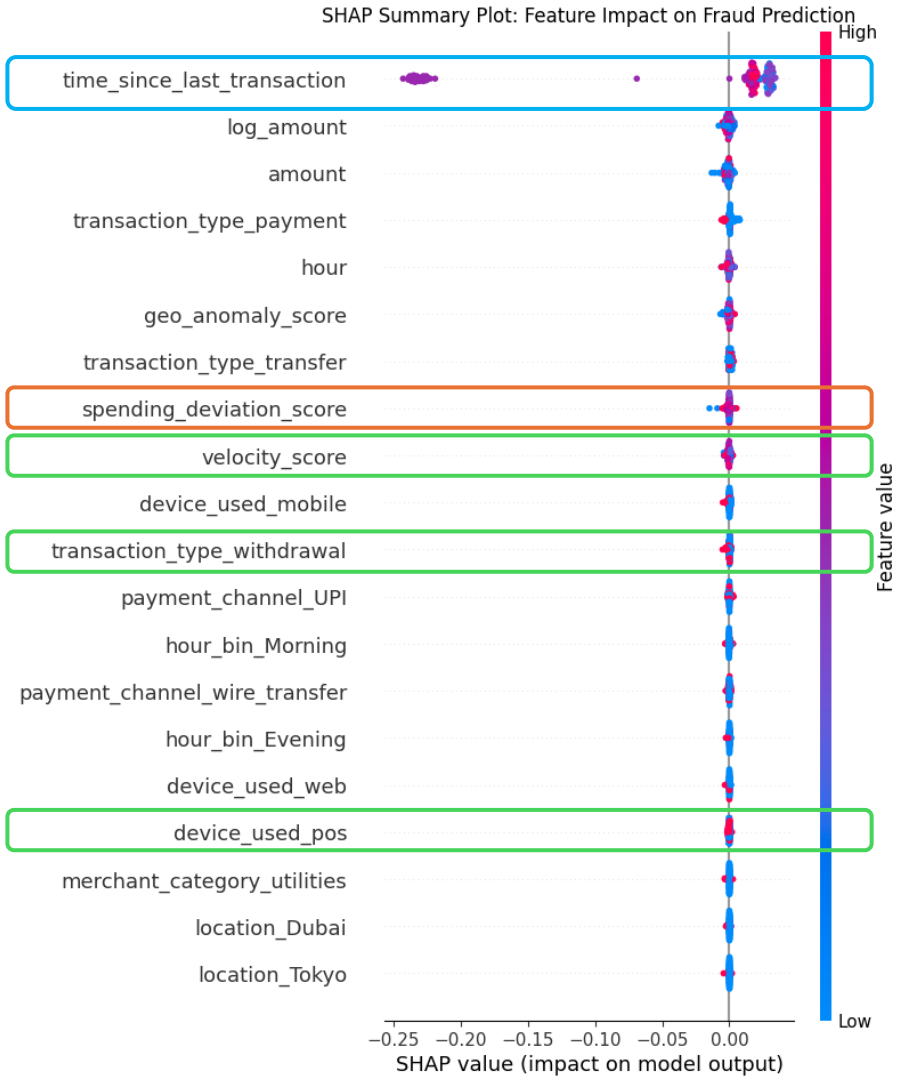


Model effectively ranks high-risk transactions well above baseline on real-world credit card transactions dataset.

Common feature: Transaction Time, Amount, Class (Fraud/Non-Fraud)

Model performance is highly dependent on the **quality and availability of behavioral signals** in the data, highlighting the importance of data quality over model complexity

Results and Insights



SHAP explains which behaviors drive the model's fraud decisions

Timing between transactions is the strongest risk signal of fraudulent transactions

Unusual spending - higher amount or volume compared to a customer's norm - also contribute to fraud predictions

Other contextual signals such as transaction type, device used for the transaction, time of day, and the number of transactions over a recent period also matter

These insights directly inform where and when to intervene without disrupting legitimate customers

Next Steps

- Continue to **monitor emerging patterns to retrain and validate the machine learning model** – as new data comes in, regularly assess if there are emerging features that can help improve fraud detection aside from time between transactions and spending behavior.

- Implement behavioral nudges and risk-based controls to **implement adaptive risk thresholds** that respond to changes in customer behavior rather than fixed transaction limits, allowing fraud controls to scale proportionally with risk.

- Periodically **reassess behavioral nudges and intervention triggers to prevent bias** and ensure continued effectiveness as customer behavior and fraud tactics change.