Detecting Fraudulent Transactions Using Machine Learning

FATIMA CAMELA RAMOS

JAN 31, 2026

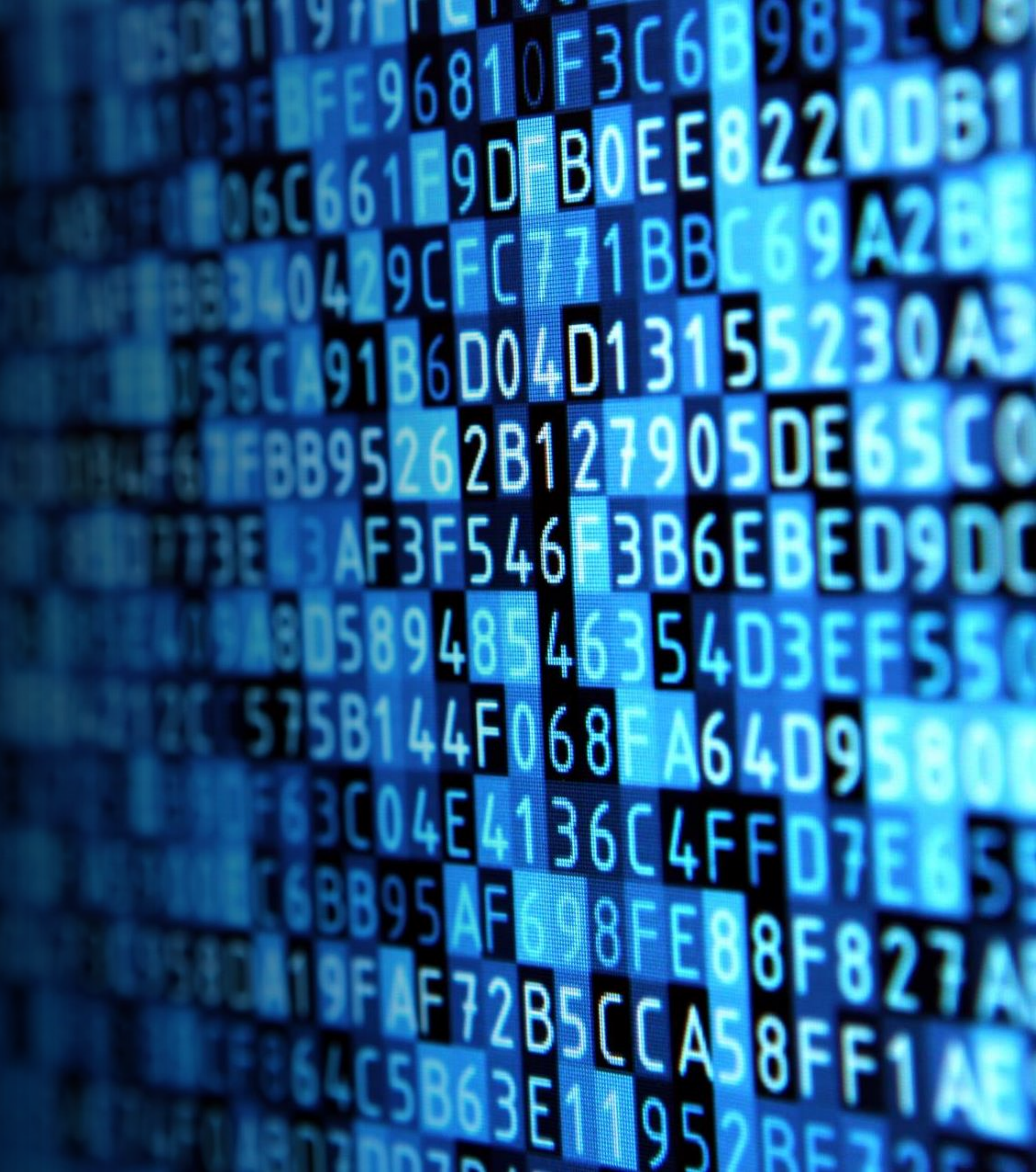
Suspected Digital Fraud Rate (~13%) in the Philippines remains high for five consecutive years

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



Fraud risk signals go beyond transaction amount and volume

Fraud tactics evolve from time to time, enabling actors to penetrate accounts and have successful transactions.

How might we be able to detect high-risk behaviors that may eventually lead to fraud?

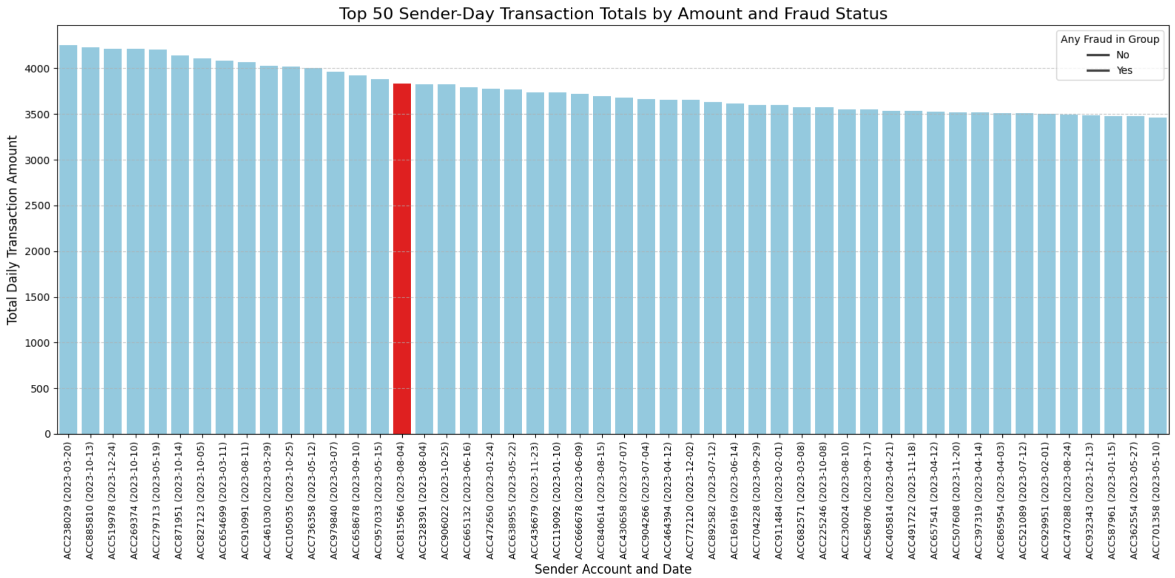


Illustration: A dataset was ranked to get the Top 50 highest transactions (amount and volume), and to label which of these are fraudulent. Only 1 out of 50 is fraud.

To detect fraudulent behaviors, transaction patterns over time will be analyzed using machine learning techniques

- Fraudulent behavior is non-linear, requiring models that can distinguish legitimate from fraudulent transactions

- Machine Learning models “learn” as new data comes in, which may improve detection strategies over time. This also allows the model to be replicable and relevant across similar use cases.

- For this use case, we will focus on building a model that can detect fraudulent activity effectively, striking a balance between fraud prevention and minimizing unnecessary transaction reviews.

Machine learning model ‘learn’ the patterns of fraudulent behavior from historical transaction data

Dataset

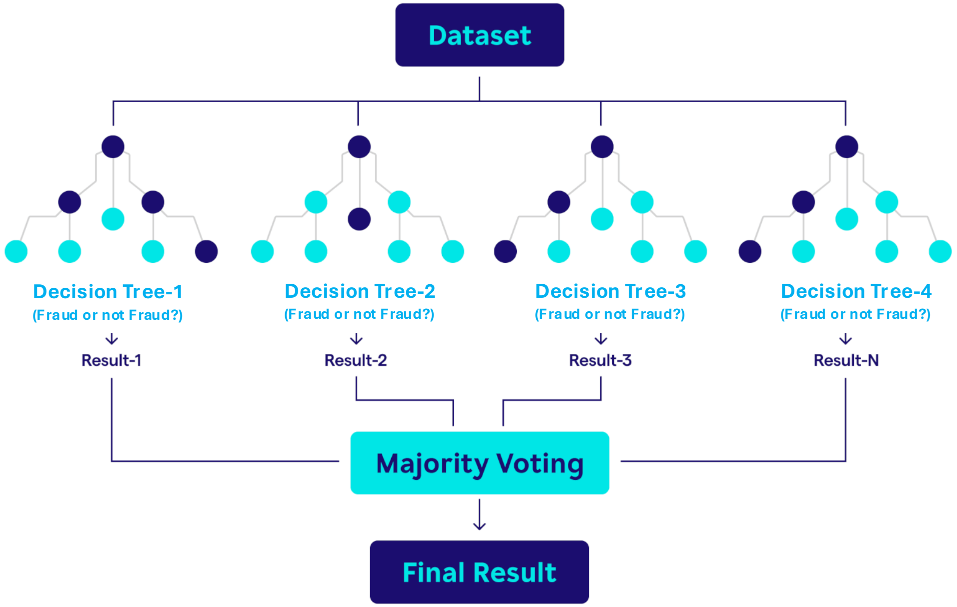
Historical transaction data is used to identify patterns associated with fraudulent behavior

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



Dataset

Decision Tree-1
(Fraud or not Fraud?)
Result-1

Decision Tree-2
(Fraud or not Fraud?)
Result-2

Decision Tree-3
(Fraud or not Fraud?)
Result-3

Decision Tree-4
(Fraud or not Fraud?)
Result-N

Majority Voting

Final Result

Model Evaluation

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

External Validation

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

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.

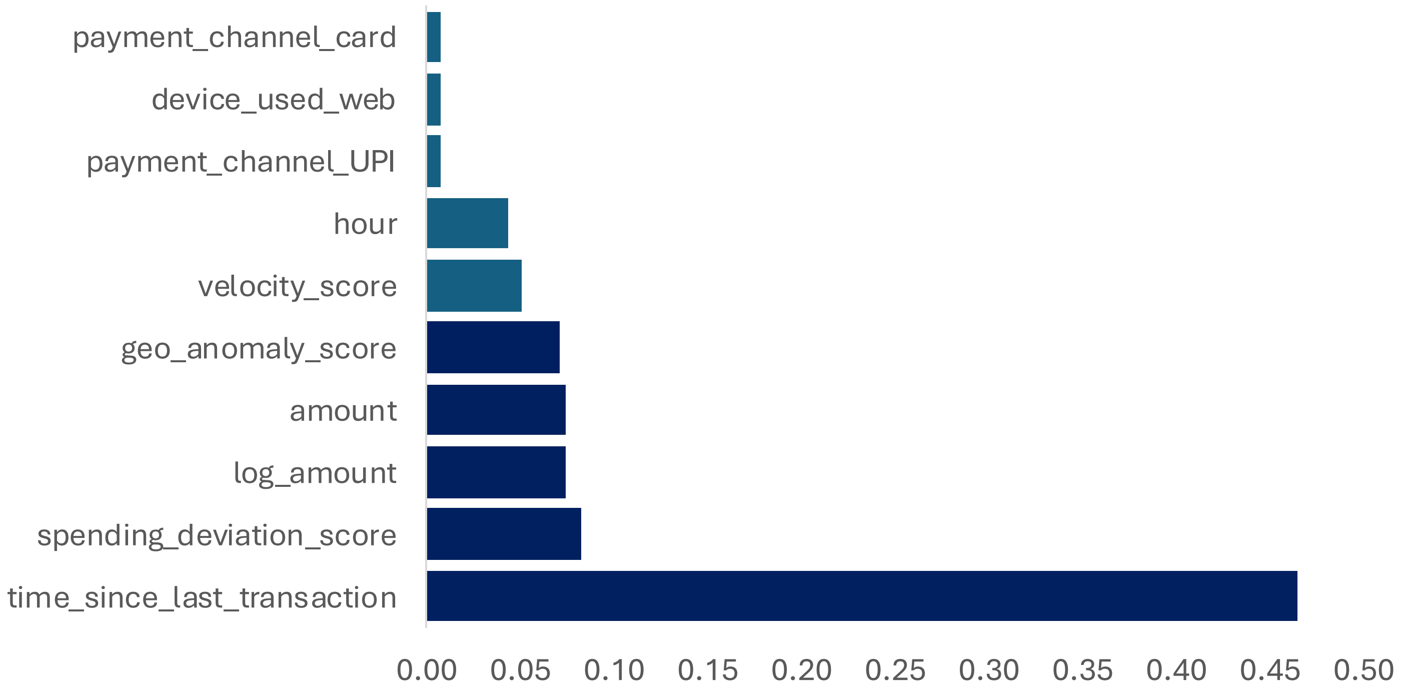
- The dataset is **highly imbalanced**, with **fraudulent transactions representing only ~3–4%** of total activity. This reflects real-world conditions and informed the use of **classification models optimized for recall and its ability to prioritize real fraud cases over low-risk transactions in a highly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

Through feature engineering, important 'signals' for model training were identified

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

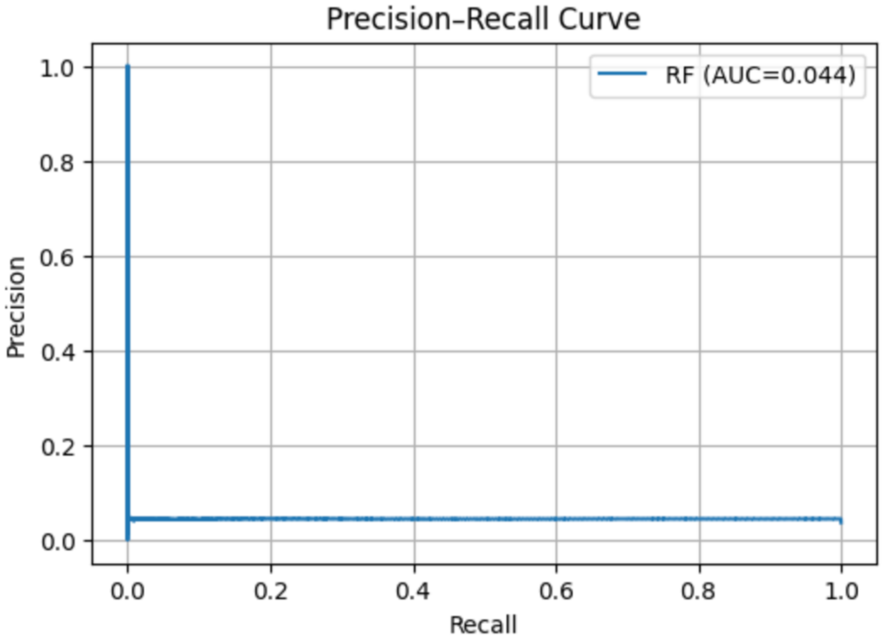


From explanatory data analysis of the dataset, the **time between transactions is the strongest indicator** of fraudulent behavior.

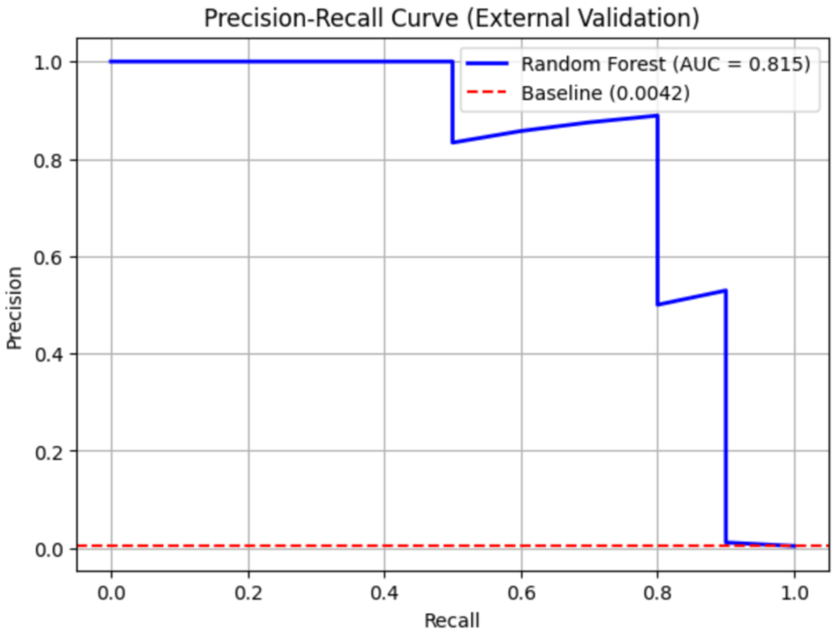
The shorter the gap between transactions, the more likely it is to be fraudulent. Other important features are:

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

Results and Insights



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

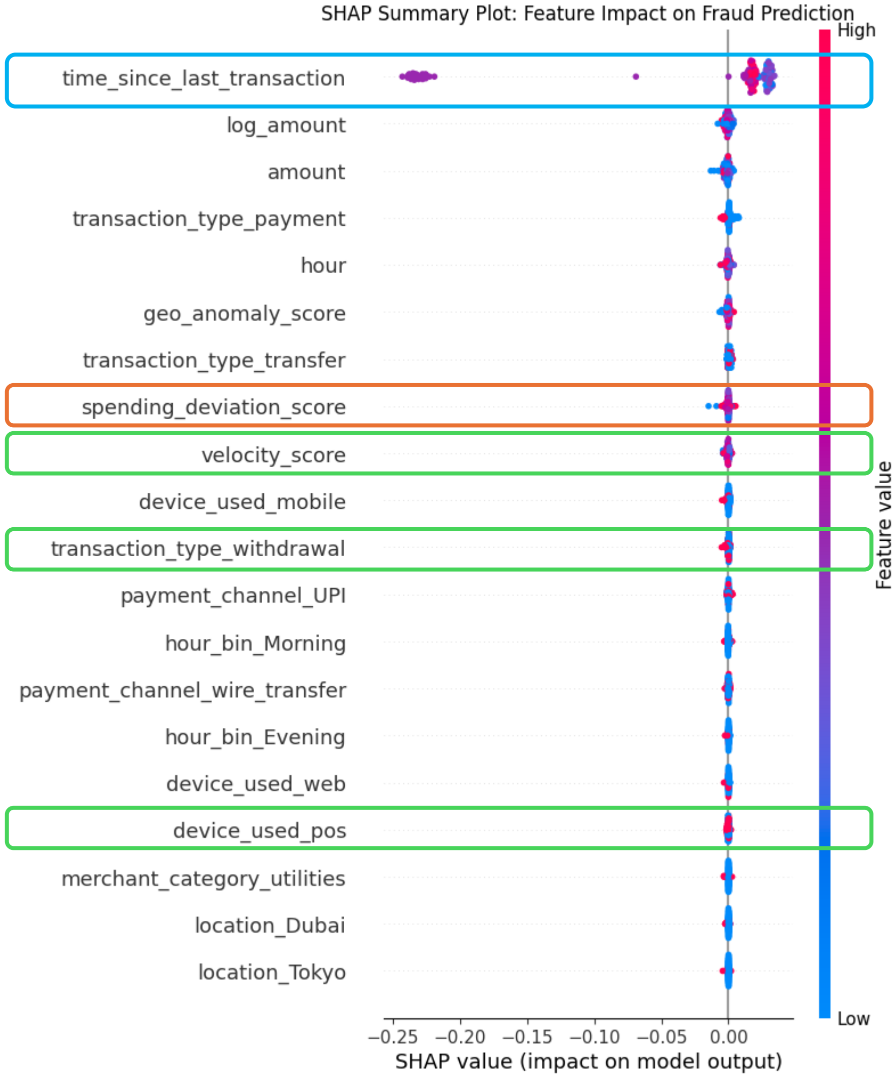


Model effectively ranks high-risk transactions well above baseline on real-world credit card transactions dataset.

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.