Supervised Fraud Detection Analysis

Human-Readable Machine Learning for Fraud Risk Scoring

Prepared by: Junior Data Scientist

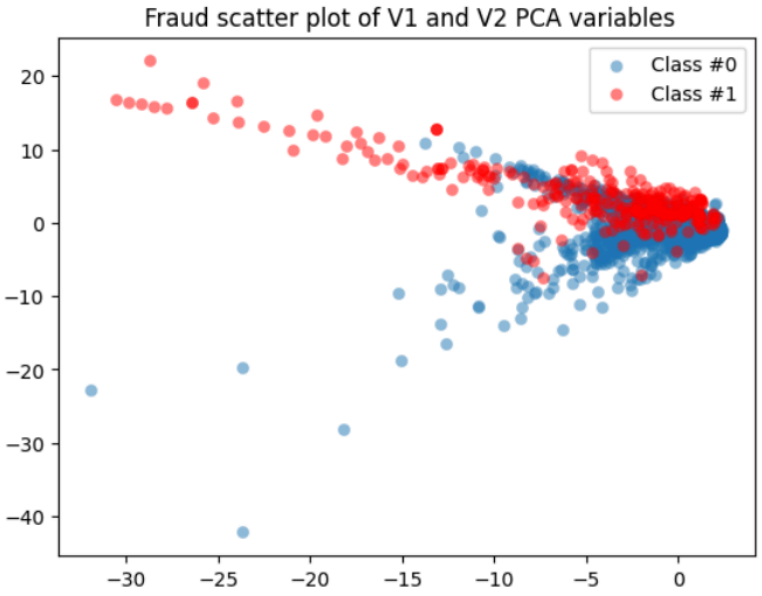
Date: July 2025

Problem Statement

This project focuses on detecting fraudulent online purchases using a supervised machine learning pipeline. Given a labeled dataset of credit card transactions, the goal was to create a high-performing and interpretable fraud detection model using Random Forest, while addressing the issue of extreme class imbalance. The dataset included anonymized (unknown) feature variables — V1 through V28 — requiring the model to learn complex patterns in hidden behavioral dimensions, without human-understandable context. SHAP was used to bring interpretability to these unknown features, making it easier to communicate model decisions to stakeholders.

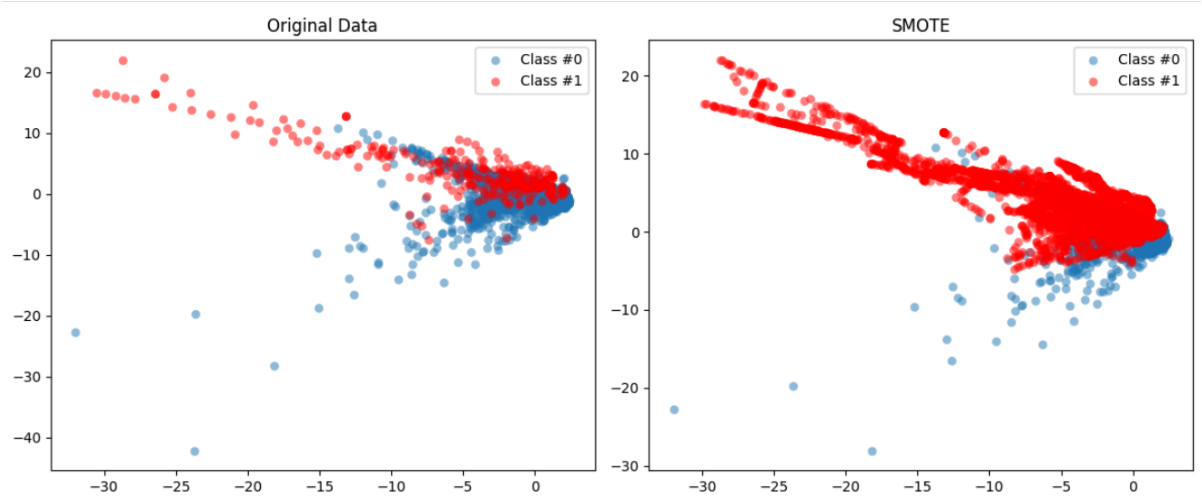
Exploratory Data Analysis

The dataset contains anonymized numerical features (V1–V28), transaction amount, and binary fraud labels. Initial exploration showed a significant class imbalance, with only 4% of the transactions marked as fraudulent, necessitating advanced sampling techniques such as SMOTE.



Handling Imbalance with SMOTE

To handle class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training data only. This generated synthetic fraud samples to balance the model's learning process without leaking information into the test set. A pipeline approach was used to prevent data leakage during cross-validation.



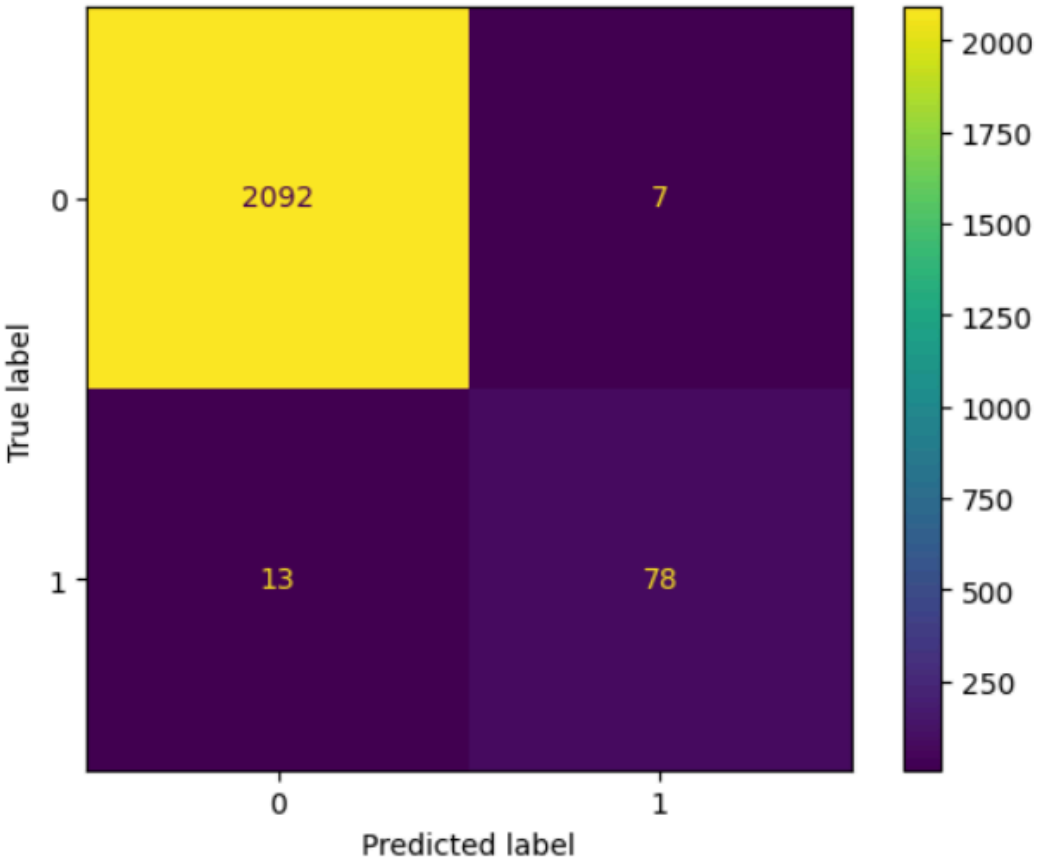
Model Building and Performance

A Random Forest classifier was trained using data balanced with SMOTE (Synthetic Minority Over-sampling Technique) to address the extreme class imbalance in the dataset. The model was evaluated using standard classification metrics, with a particular focus on recall due to the high cost of missing fraudulent cases.

The model demonstrated strong performance:

- **Accuracy:** 99%
- **Precision:** 92%
- **Recall:** 86%
- **F1 Score:** 89%
- **ROC AUC:** 98%

These results indicate a well-calibrated model capable of identifying fraudulent transactions with high precision while maintaining a strong ability to catch most fraud cases, which is critical in high-stakes financial settings.

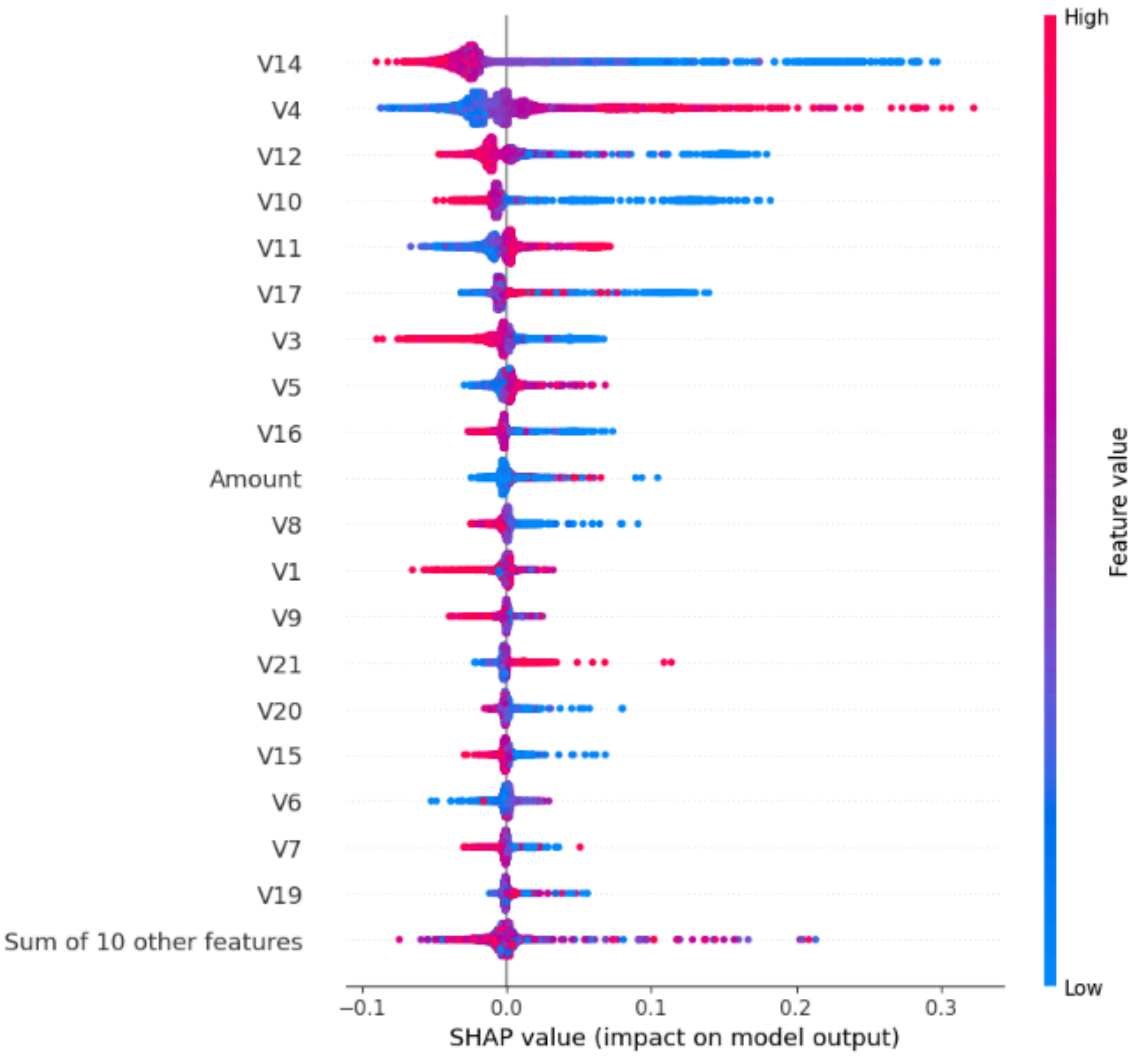


	precision	recall	f1-score	support
Not Fraud	0.99	1.00	1.00	2099
Fraud	0.92	0.86	0.89	91
accuracy			0.99	2190
macro avg	0.96	0.93	0.94	2190
weighted avg	0.99	0.99	0.99	2190

Model Interpretability (XAI)

SHAP values were used to interpret the Random Forest model, enabling a detailed understanding of how each feature influenced individual fraud predictions. Beeswarm plots revealed which anonymized features (e.g., V1–V28) most strongly contributed to a transaction being flagged as fraudulent.

Despite the variables being anonymized, SHAP provided transparency by highlighting relative importance and directionality. In real-world deployment, this interpretability ensures that authorized analysts or compliance teams—those with access to the underlying feature definitions—can trace decisions back to their root causes, supporting accountability, auditability, and trust in the model's outputs.



Final Insights

This project demonstrates a robust end-to-end fraud detection pipeline: from preprocessing and SMOTE balancing to model evaluation and explainability. The SHAP-enhanced interpretability ensures that the model is not a black box, supporting transparent, data-driven fraud prevention