CREDIT CARD FRAUD DETECTION

Kris Rubiano

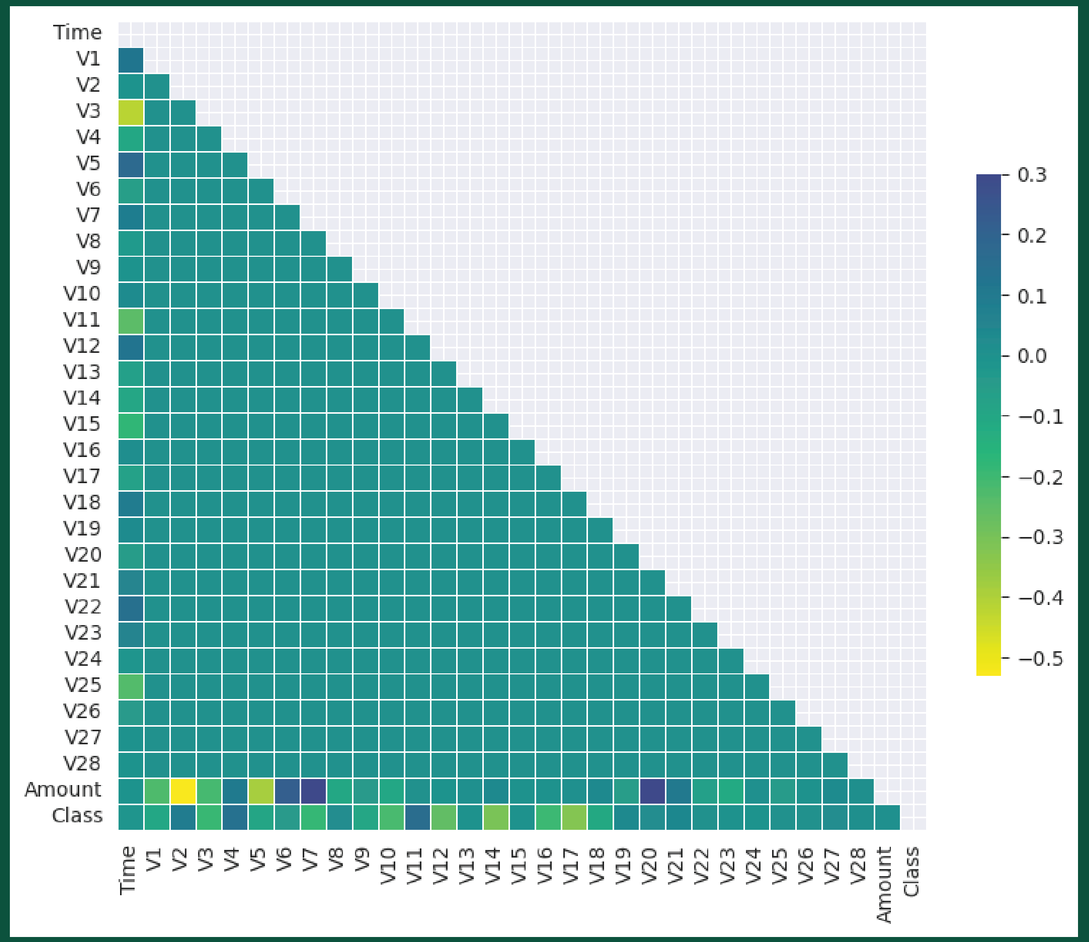
OBJECTIVE

Build an optimized prediction and anomaly detection model for fraudulent credit card transactions.

DATA

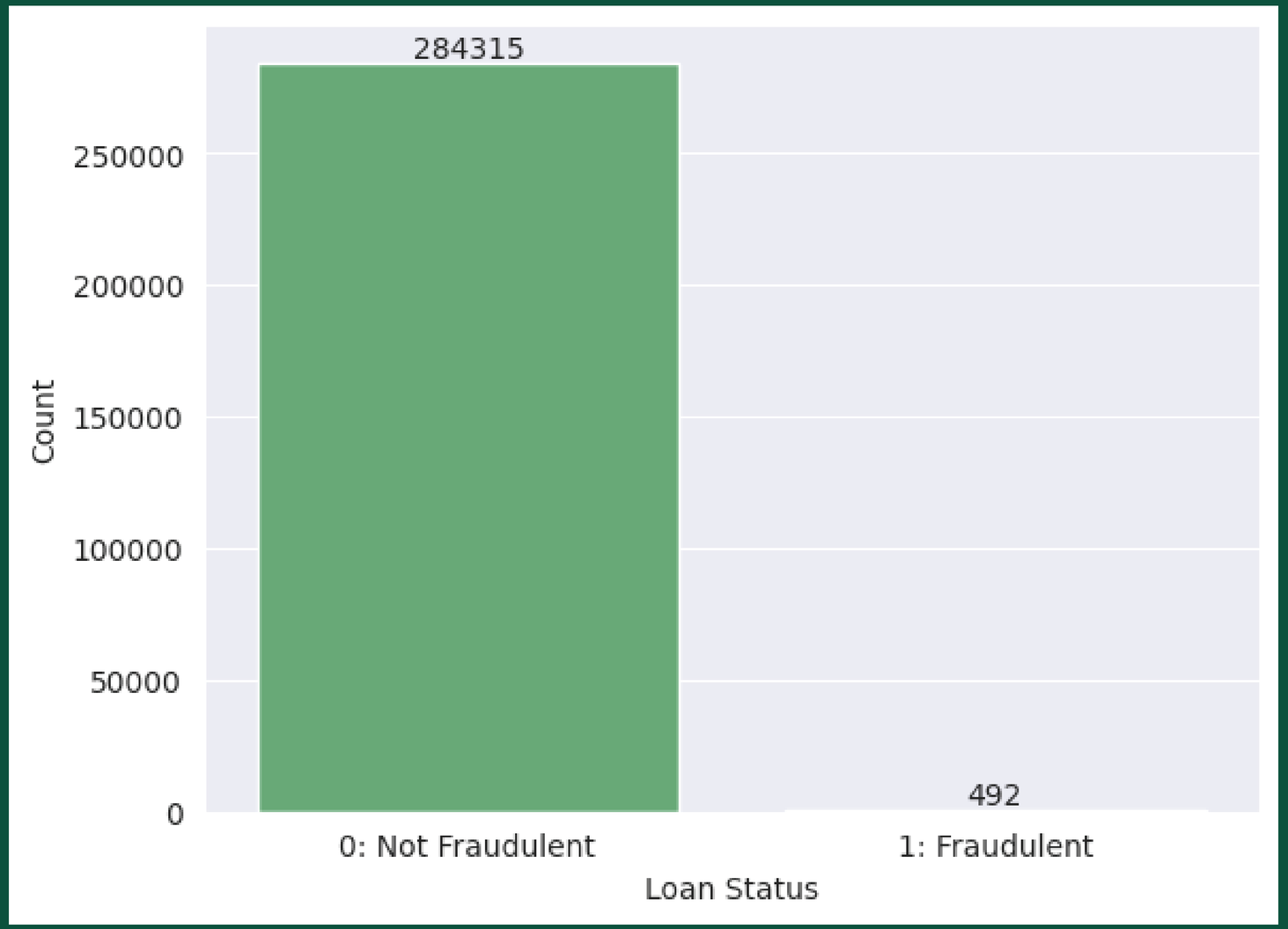
This dataset contains credit card transactions within 2 days in September by European cardholders.

It was collected and analyzed during a research collaboration of Worldline and the Machine Learning Group of Université Libre de Bruxelles.



IMBALANCE

In this dataset, there are only 492 fraudulent cases! This means, about 99.8% of the data are legitimate transactions, and only about 0.2% is fraudulent.



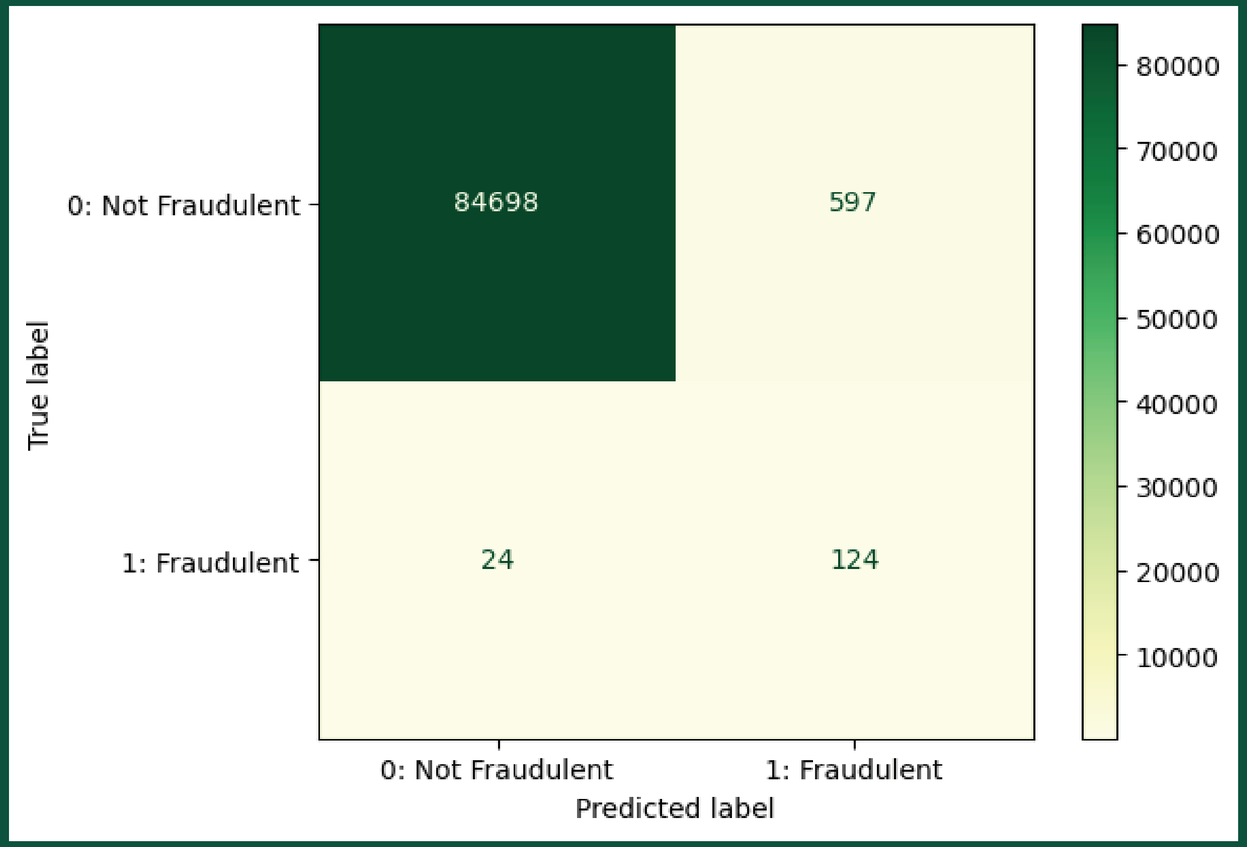
PREDICTIVE MODELS

- Logistic Regression
- Gradient Boosting
- Decision Tree
- Neural Network
- Random Forest

PREDICTIVE MODELS

- Logistic Regression
- Gradient Boosting
- Decision Tree
- Neural Network
- Random Forest

Logistic Regression

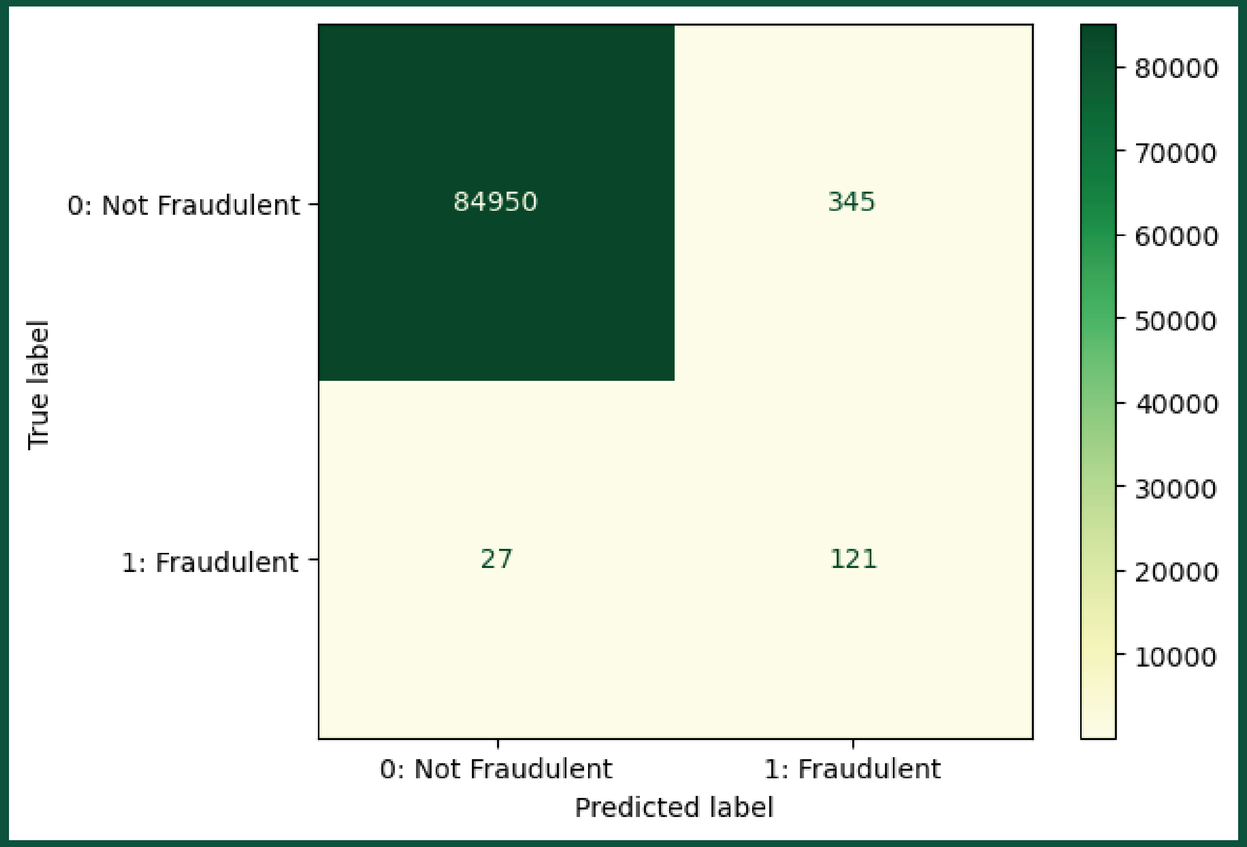


	precision	recall	f1-score	support
0	1.00	0.99	1.00	85295
1	0.17	0.84	0.29	148
accuracy			0.99	85443
macro avg	0.59	0.92	0.64	85443
weighted avg	1.00	0.99	1.00	85443

- Overall accuracy is 99%.
- Class 1 recall is at 84%.

- Fastest model.

Neural Network



	precision	recall	f1-score	support
0	1.00	1.00	1.00	85295
1	0.26	0.82	0.39	148
accuracy			1.00	85443
macro avg	0.63	0.91	0.70	85443
weighted avg	1.00	1.00	1.00	85443

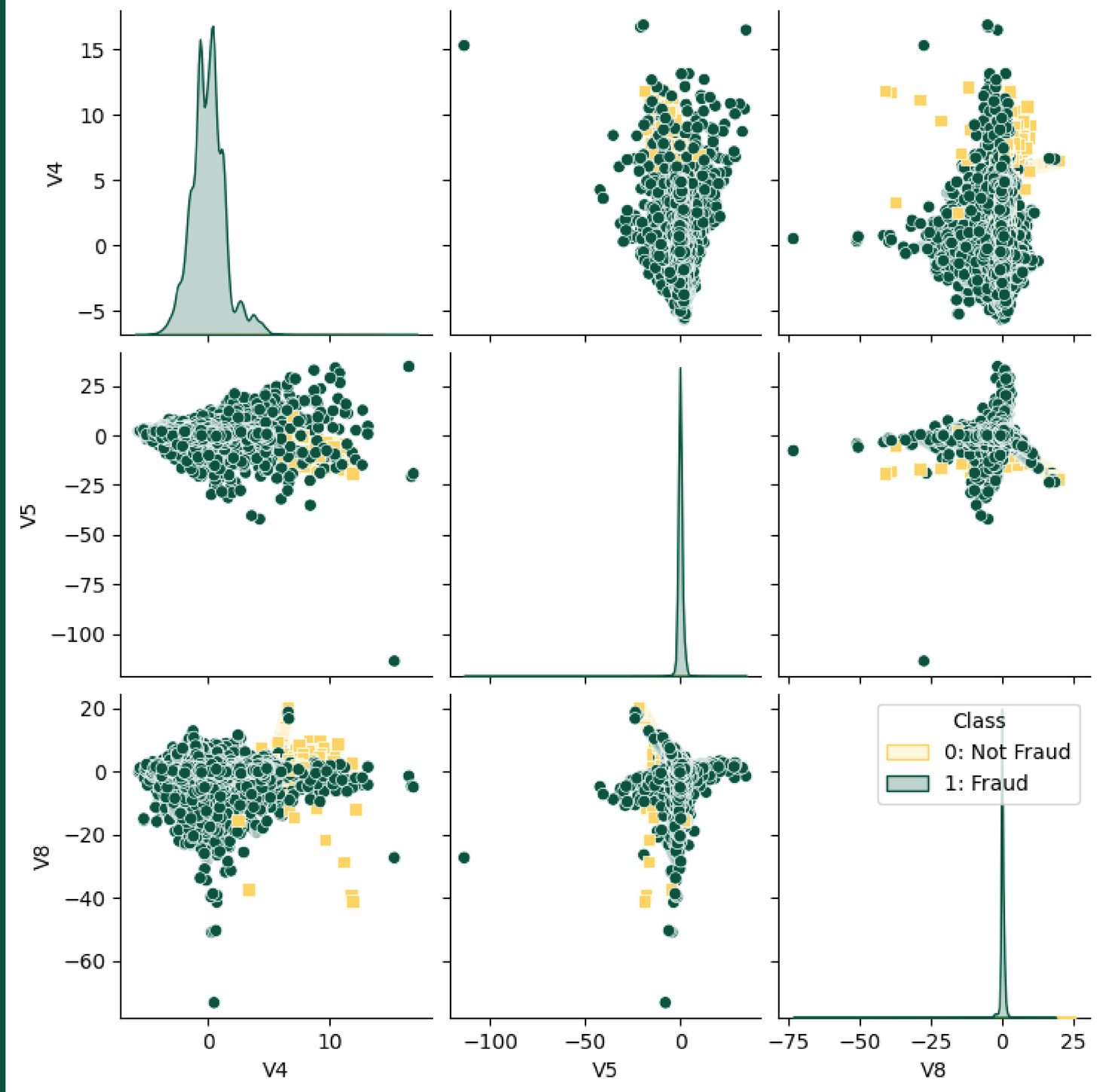
- Overall accuracy is 1% higher than Logistic Regression's model.
- Class 1 recall is at 82%.
- Very similar report to a tuned Gradient Boosting model, but arrives at these results much faster.

IMPORTANT FEATURES

using recursive feature elimination

with cross-validation

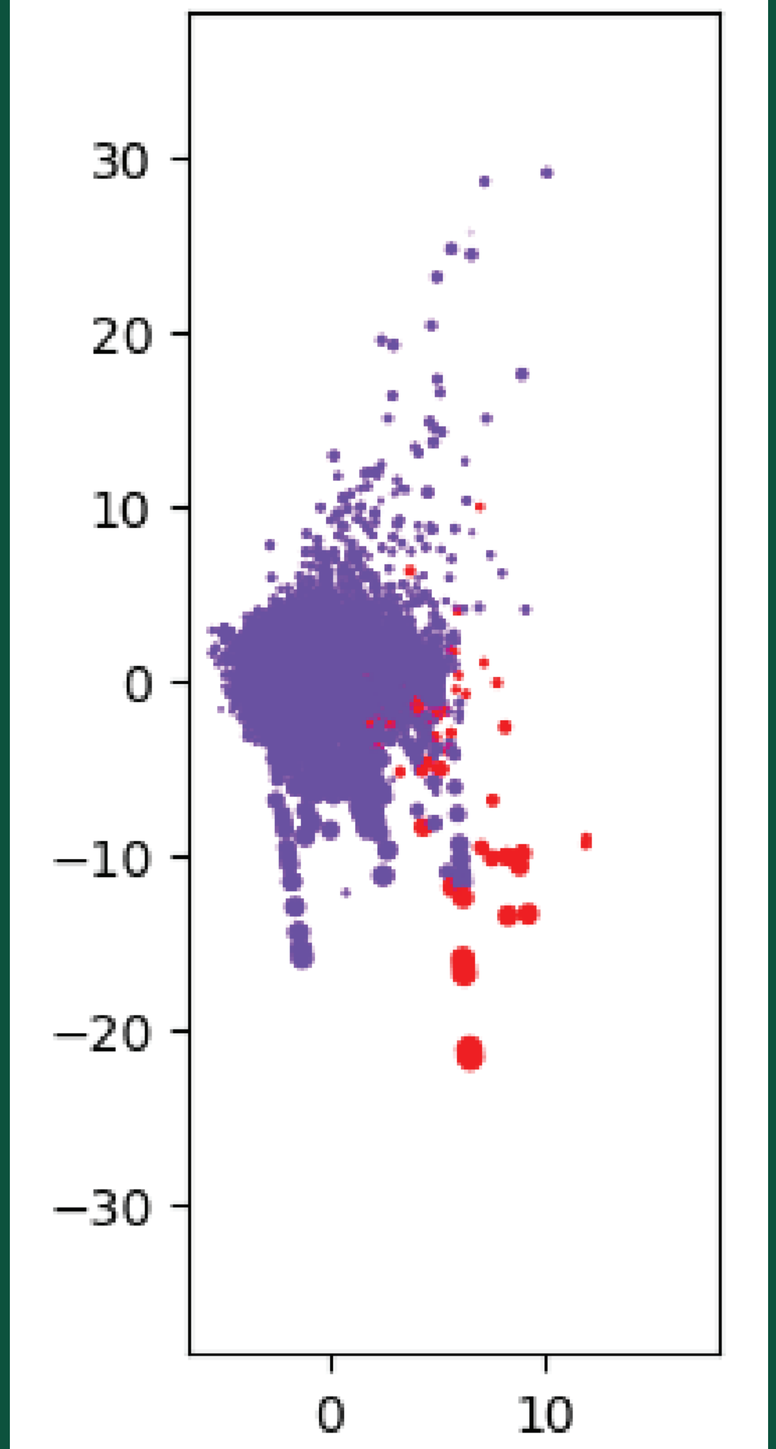
1. V4
2. V5
3. V8



ANOMALIES

using SVM

Baseline
Model



NEXT STEPS

- Continue to tune both best predictive models to improve recall scores.

- Improve anomaly detection models.

Thanks!

Additional

Photos: Unsplash, The Noun Project

“An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes”
https://www.sciencedirect.com/science/article/pii/S2214212623002028