Banking on Fraud

Using Machine Learning to minimize the toll fraud takes on banks

Matthew Kwee

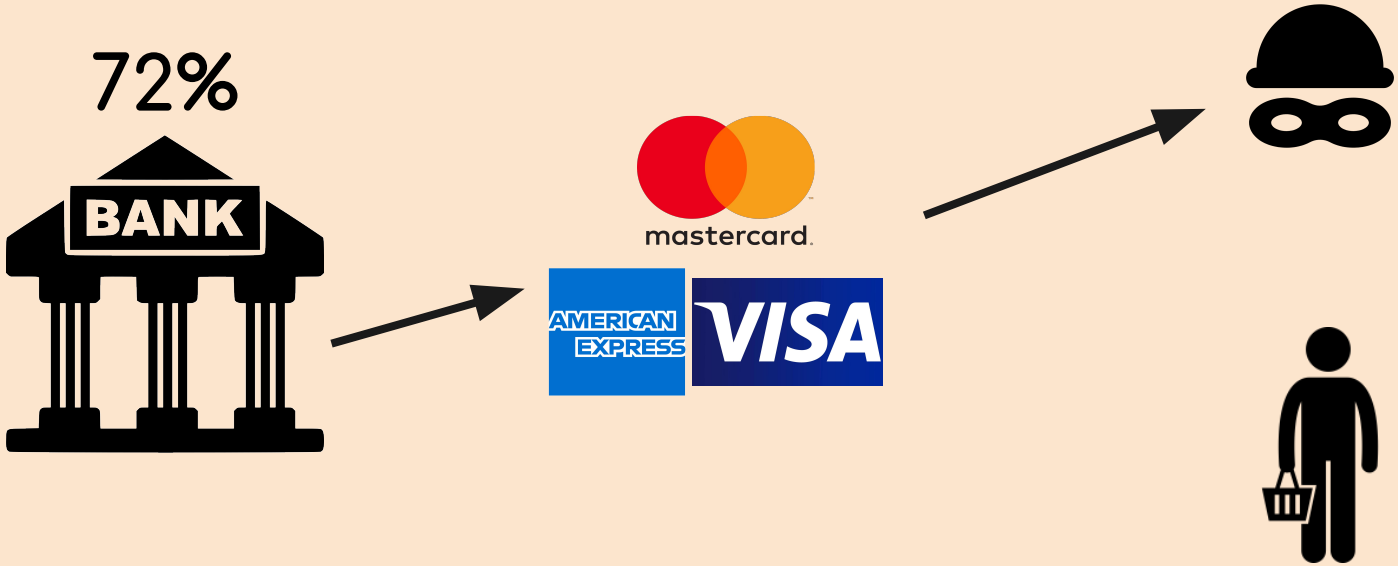
29 October 2021

Fraud is a problem for everyone

In 2018, $24.26 Billion was lost due to payment card fraud worldwide

https://shiftprocessing.com/credit-card-fraud-statistics/

When fraud occurs, banks are often on the hook



Investigating fraud is expensive!

Forensic projects cost upwards of $8 per case.

With over 1 billion transactions worldwide every day, we can’t afford to audit every transaction.

https://www.valid8financial.com/post/bank-fraud-investigation-cost

Objective:

Minimize cost of fraud for banks.

Data

Fraud Detection Dataset[1]:

- 284,807 datapoints
- 30 features (28 composite)
- 492 positive cases

Tools

Modeling: SciKit-Learn and XGBoost

Data handling: Pandas and NumPy

Visualization: Google Sheets

1. https://www.kaggle.com/mlg-ulb/creditcardfraud

Assumptions

- Customers do not churn because of false positives or negatives.[1]
- A fraudulent transaction costs the bank about €85[2]
- Investigating a potentially fraudulent transaction costs about €6[3]*

*The current USD-EUR exchange rate is approximately $1.17 to €1.00

1. Model has built-in churn costs; all we have to do is re-fit the model when we have data
2. https://www.ecb.europa.eu/pub/pdf/other/4th_card_fraud_report.en.pdf
3. https://www.valid8financial.com/post/bank-fraud-investigation-cost

Metrics

$$C = 6(P) + 85(F_{neg})$$

Naive Models:

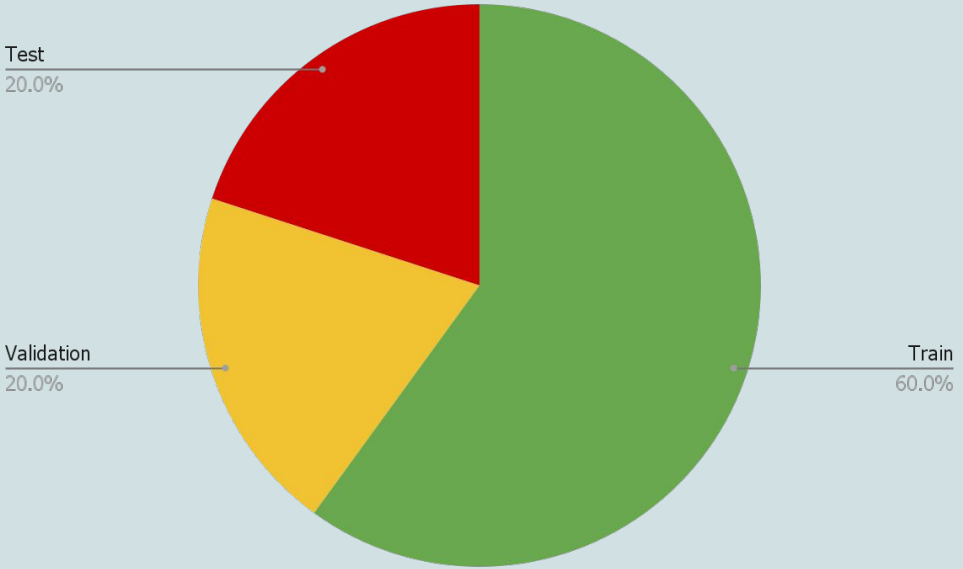
Cost/Transaction (Legitimate): €0.146584

Cost/Transaction (Fraudulent): €6.00

Modeling

Models evaluated:

- k-Nearest-Neighbors
- Logistic Regression
- Random Forest
- Gradient-Boosted Trees



Techniques utilized

- Adjusting probability threshold
- Undersampling negative cases
- Increasing positive cases' weight
- #Features Selected

XGBoost-exclusive metrics:

- Learning Rate
- Child Weight
- Sub-sampling

Models

- Logistic Regression
- Random Forest
- Gradient-Boosted Trees

Fraud detection models are often “black-box,” so a high recall score (and low cost/transaction) is prioritized over interpretability.

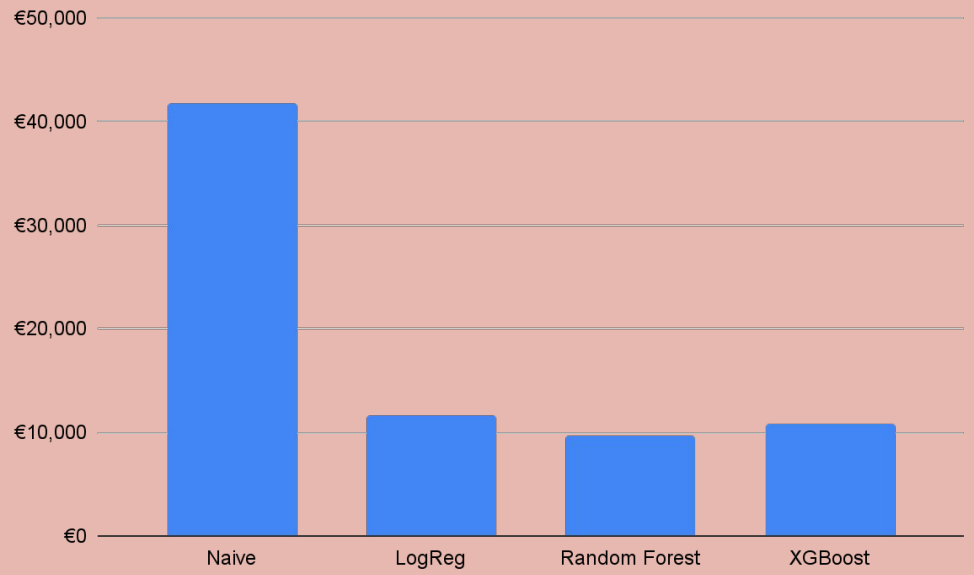
	Logistic Regression	Random Forest	GB Trees
Accuracy	0.998683	0.999719	0.999034
Precision	0.630769	0.918919	0.704545
Recall	0.752294	0.935780	0.853211
F1	0.686192	0.927273	0.771784
Cost/Transaction	€0.0409889	€0.0339465	€0.0377385

Models

- Logistic Regression
- Random Forest
- Gradient-Boosted Trees

The naive baseline loses €40,000 over two days described by the dataset.

By comparison, the Random Forest model saves over €30,000 = $35000.



Future Work

Additional factors to take into account:

- Customer churn
- Purchase location
- Customer demographic

More Data!

Questions?