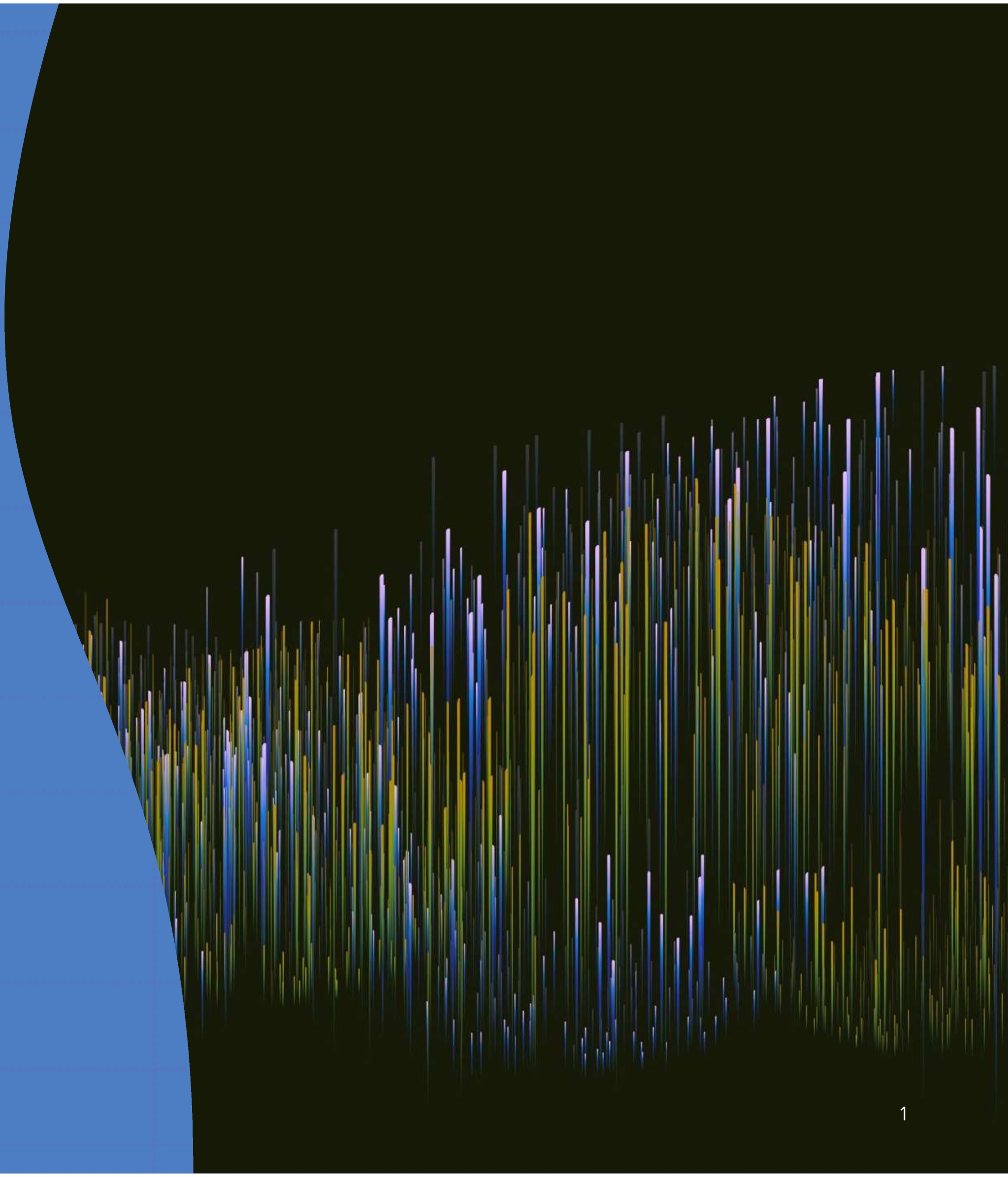
Bernard Wittmaack

Detecting Fraud in Mobile Money Transactions



Mobile Money Services (MPS)

- Quickly gaining popularity
- Heavily relied upon in nations with poor banking penetration for everyday commerce
- Vulnerable to fraud

Real MPS Data to Train Fraud Detection Models Not Readily Available

- MPS data contains personal and sensitive information and, hence, is largely unavailable outside the MPS company.

- Fraud detection researchers need data to drive their work

- Researchers developed a high-fidelity synthetic dataset based on real logs of an African MPS company.

PaySim: A Financial Mobile Money Simulator for Fraud Detection; https://www.researchgate.net/publication/313138956_PAYSIM_A_FINANCIAL_MOBILE_MONEY_SIMULATOR_FOR_FRAUD_DETECTION

Synthetic Financial Datasets For Fraud Detection; https://www.kaggle.com/ealaxi/paysim1

Executive Summary

- Three machine learning algorithms (logistic regression, random forest, and support vector machine (SVM) classifier) trained to detect fraud from synthetic data.

- Best predictive model (**SVM**) is **90% accurate**, and the precision and recall were essentially equal (**F1-score ~ 0.90**). Random forest also good.

Feature Correlation with Fraud

Positively Correlated

- Time of Transaction*
- Transfer Transaction
- Account Emptied
- Transaction Amount

*Not generalizable

Negatively Correlated

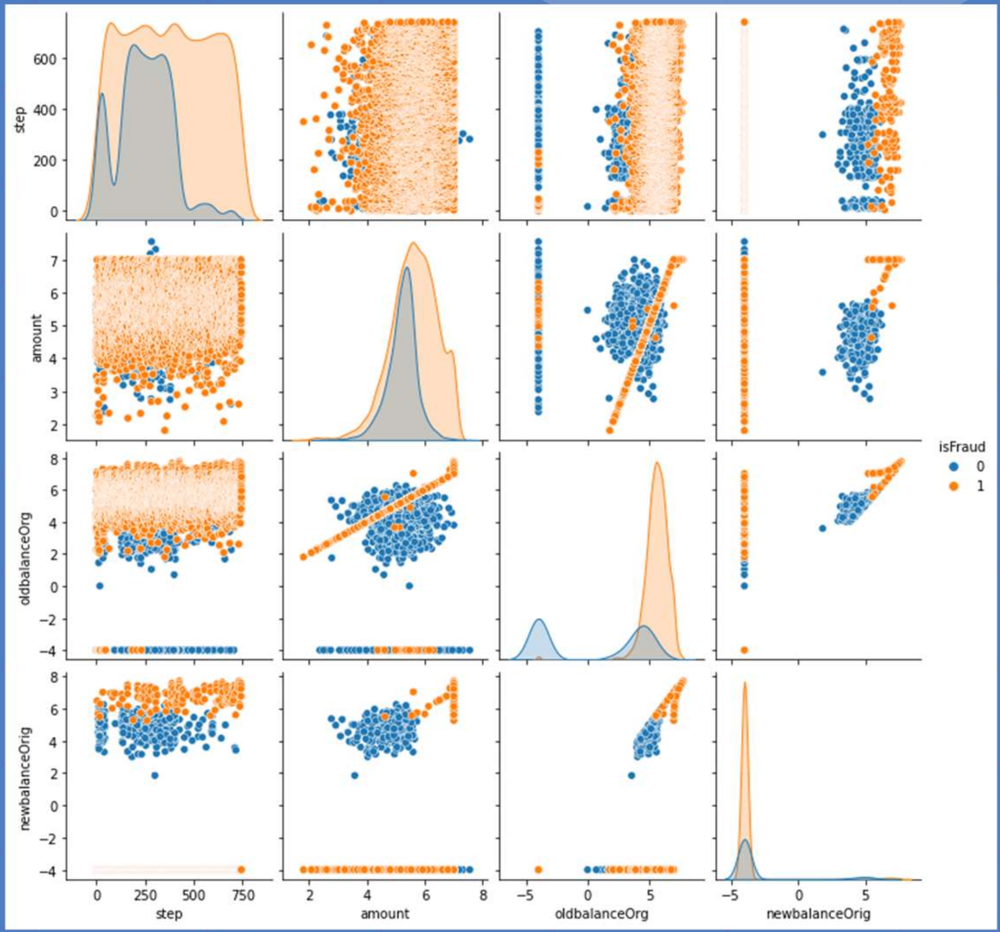
- Cash Out Transaction

Features for Detecting Fraud

Numerical Features

1. Step
2. Transaction Amount

Note the strong linear correlation between balance and amount for fraud.



Feature Engineering

- Fraudulent transactions frequently empty an account
- Create feature for whether account was emptied during transaction.

Features for Detecting Fraud

- Only 2 of the transaction types (transfer and cash out) associated with fraud.
- Transforming categorical features needed for most machine learning algorithms.

Categorical Features

1. Emptied account
2. Transaction Type

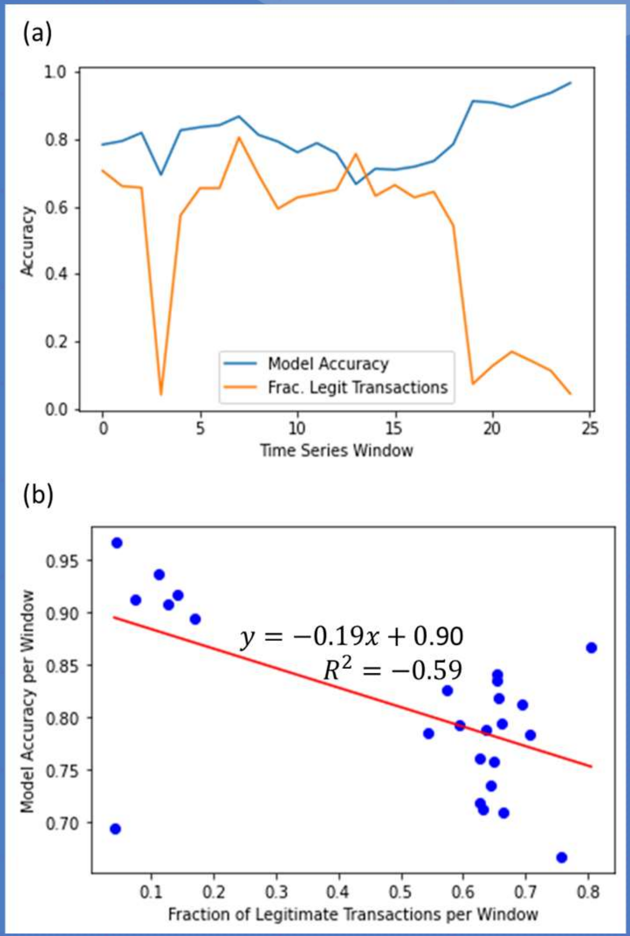
→

Transformation

Low cardinality +
Nominal =
One-hot encoding

Relevance of the Step Feature

- No reference point for step. No clear trends that would allow us to tie it to real world seasonality (e.g., weekends, day/night, holidays).
- Step is important for fraud classification in dataset but is not easily generalizable to unseen data.
- No evidence of non-stationarity.

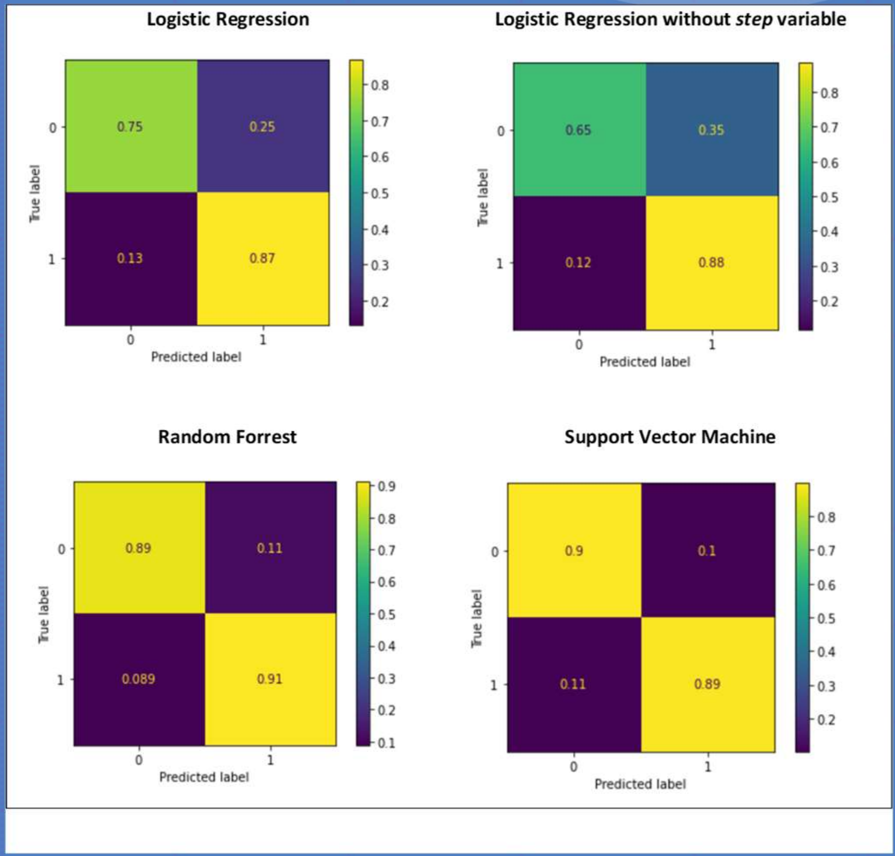


Comparison of Algorithm Performance

- All models trained with 5-fold cross-validation and validated against separate test data.

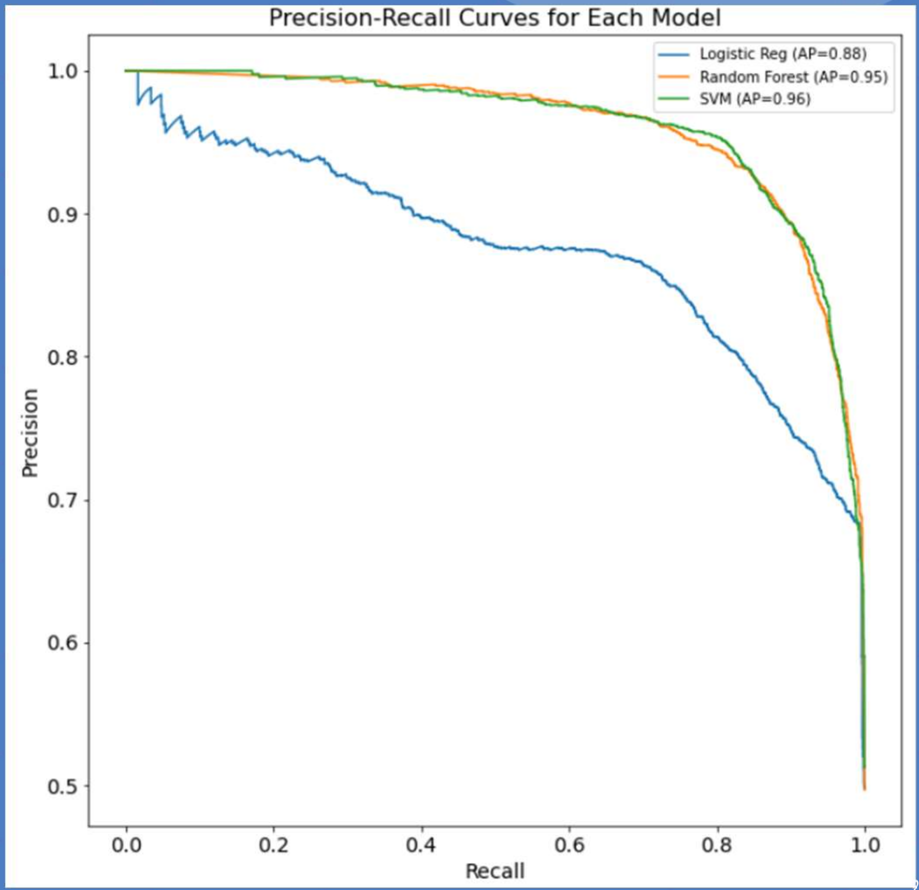
- Hyperparameter tunning led to minimal gains in F1-scores in logistic and random forest models but significant gains for SVM.

- SVM and random forest models much better at correctly classifying legitimate transactions than logistic regression.



Judging Model Performance for Fraud

- Fraud is costly and tarnishes the MPS company's reputation.
- Better to over-predict fraud as opposed to under-predict it.
- However, severe over prediction can lead to inconvenient delays for legitimate transactions.
- Can create custom loss-function to penalize false negatives.



What Does Each Feature Say About Fraud

Feature Correlation with Fraud

Positively Correlated

- Time of Transaction*
- Transfer Transaction
- Account Emptied
- Transaction Amount

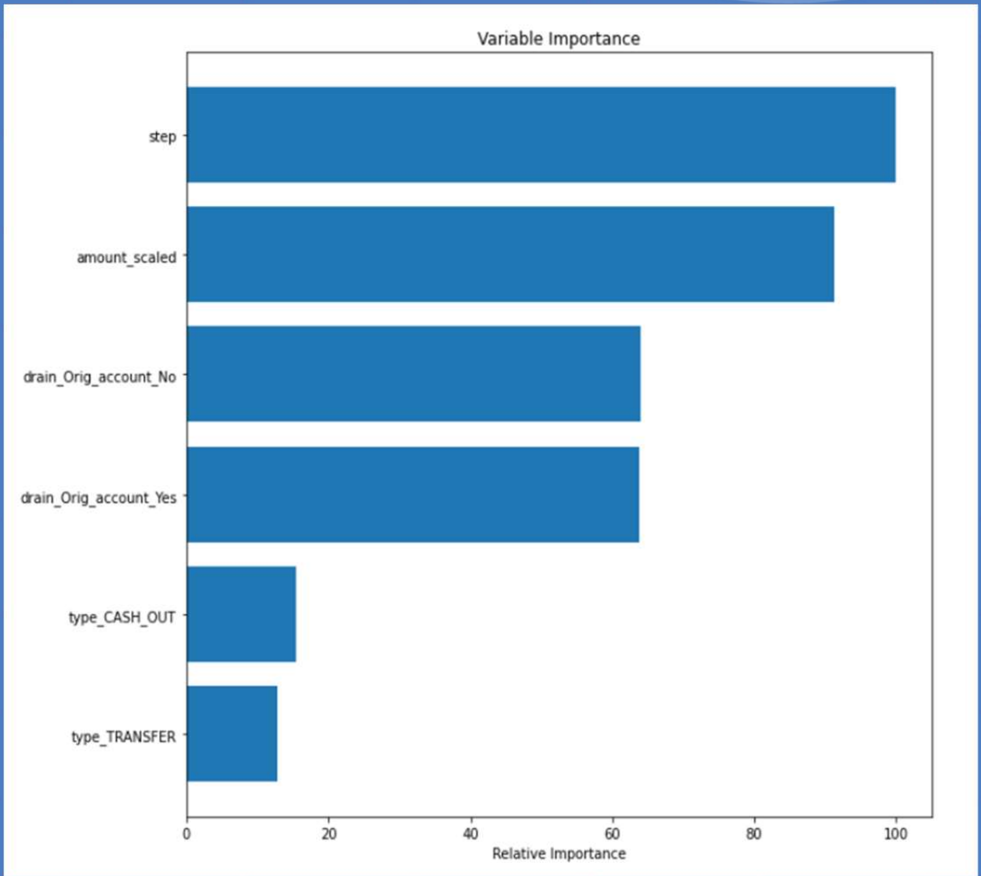
*Not generalizable

Negatively Correlated

- Cash Out Transaction

- All logistic regression coefficients statistically significant ($p<0.001$)
- Gini importance suggests step and amount are top features.

Relative Gini Importance from Random Forest Model



Summary and Future Work

- **SVM** and **random forest** models perform similarly well (f1=0.9) and better than logistic regression (f1 = 0.8).
- Use logistic model to better understand interactions between features.
- Need to better understand time component for generalization.