Fraud Detection

Problem

Fraud in online transactions:

- Loss of revenue
- Decreased customer satisfaction

Solution

Fraud detection system:

- Accurate fraud prediction
- Less false alarms

Data

- IEEE Computational Intelligence Society Fraud Detection
- Vesta Corporation's real-world e-commerce transactions
- 590,540 observations, 434 variables, 118 features

Workflow

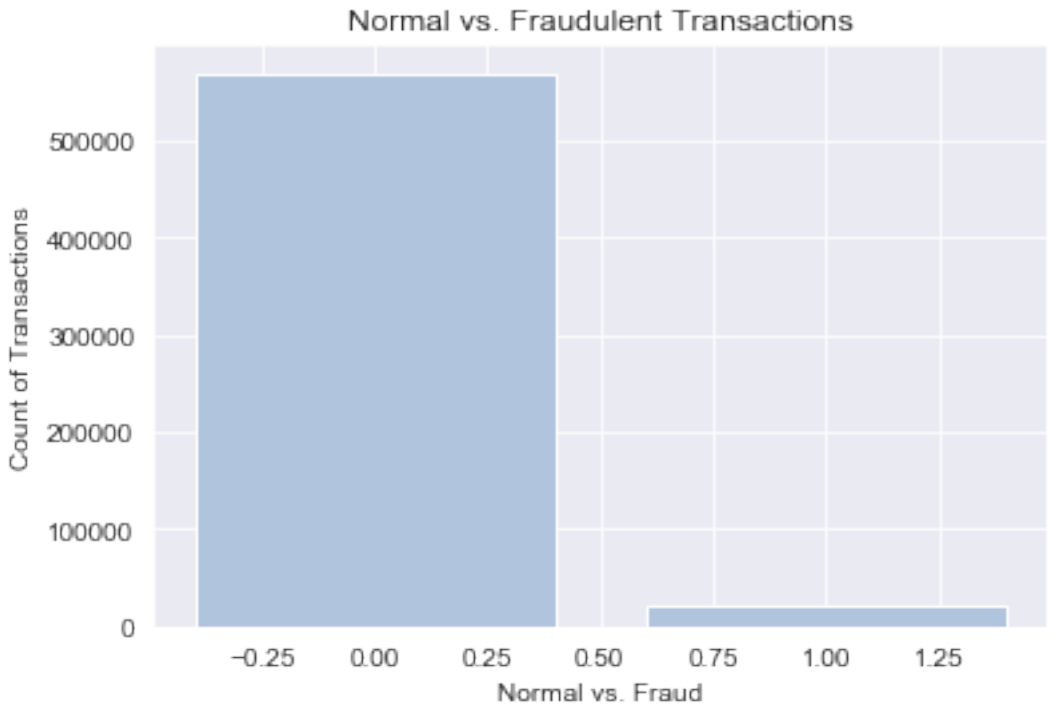
I. Data Preparation:

- Data Cleaning
- Exploratory Data Analysis
- Feature Engineering

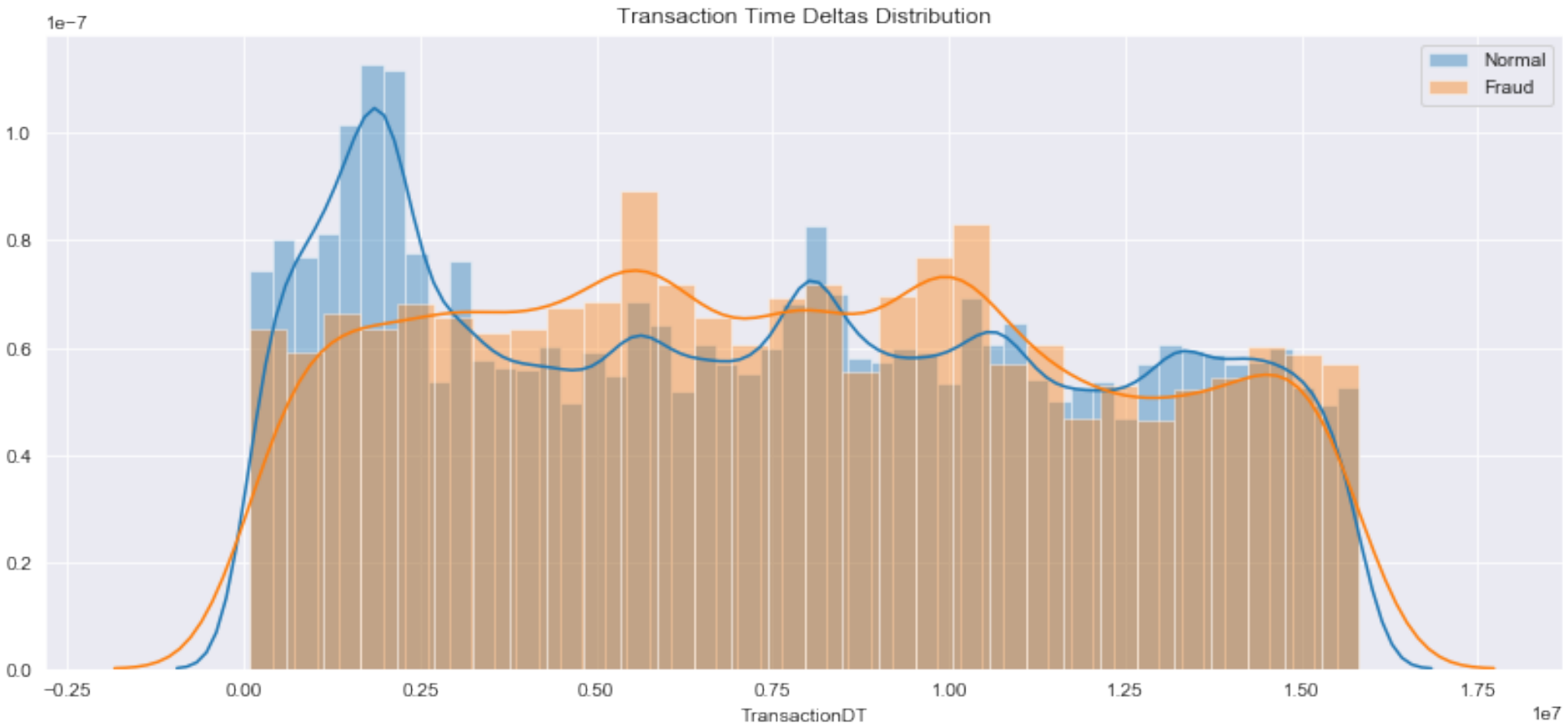
II. Modeling:

- Class Imbalance
- Model Evaluation and Selection
- Hyperparameter Tuning

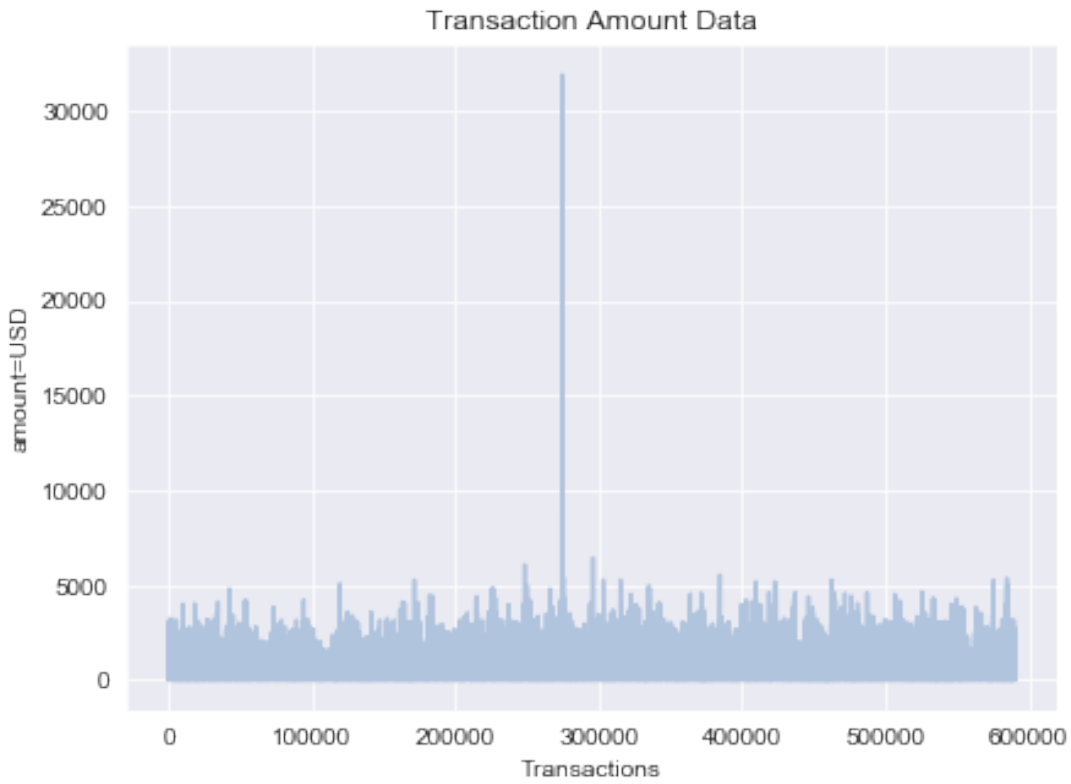
Target Variable: class imbalance

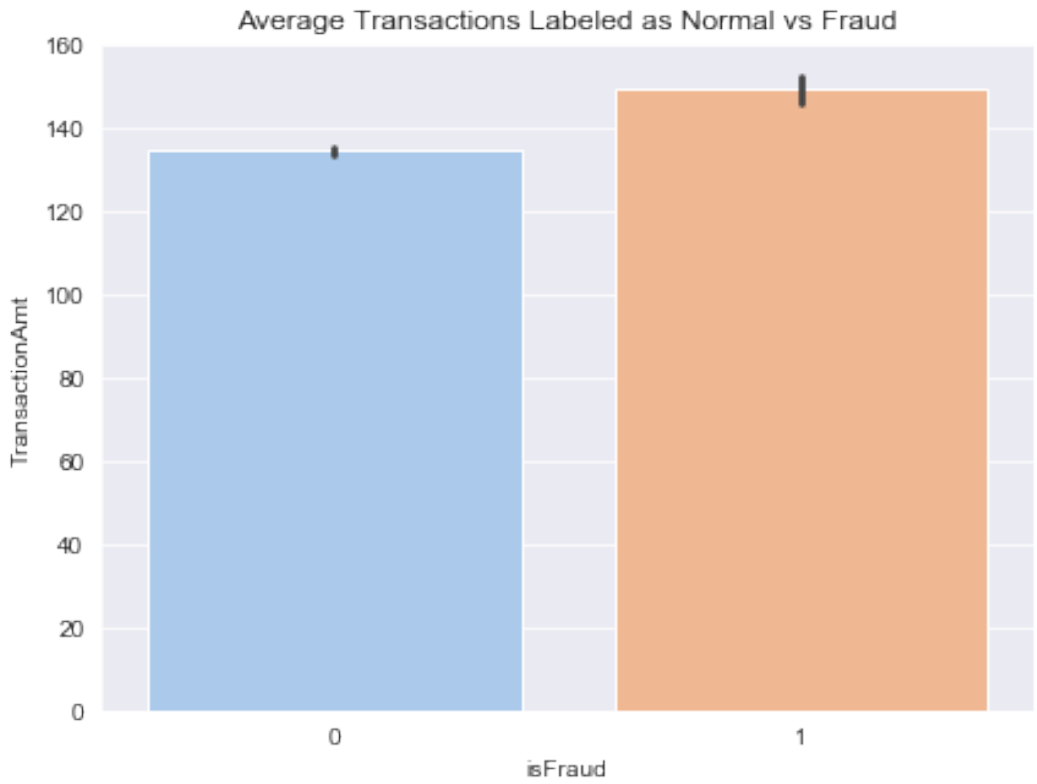


Transaction Time Deltas

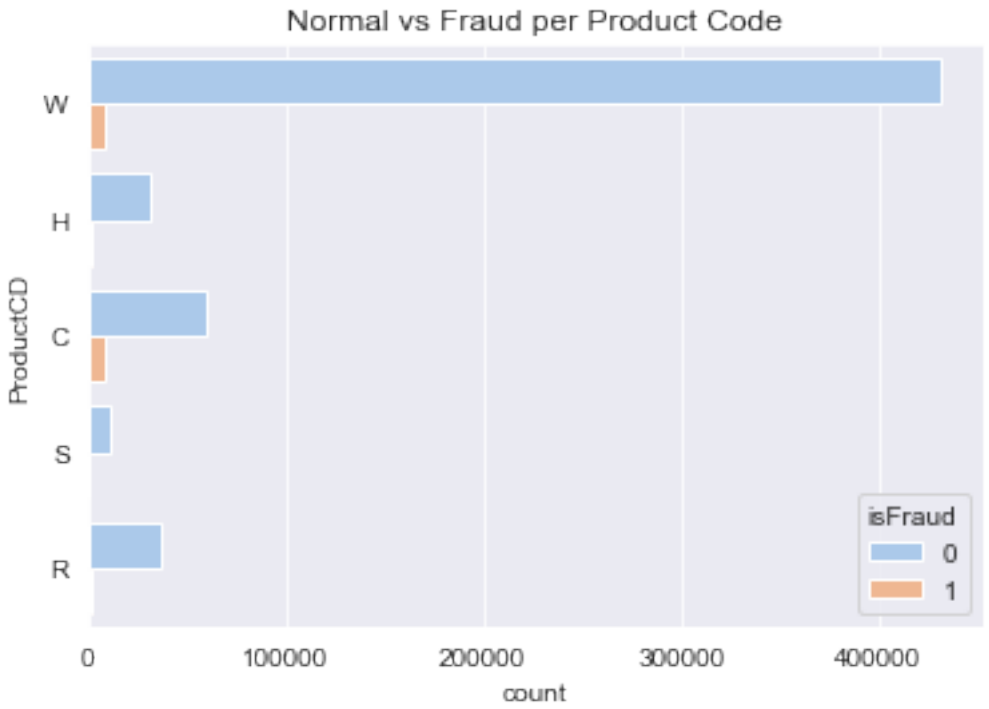


Transaction Amount

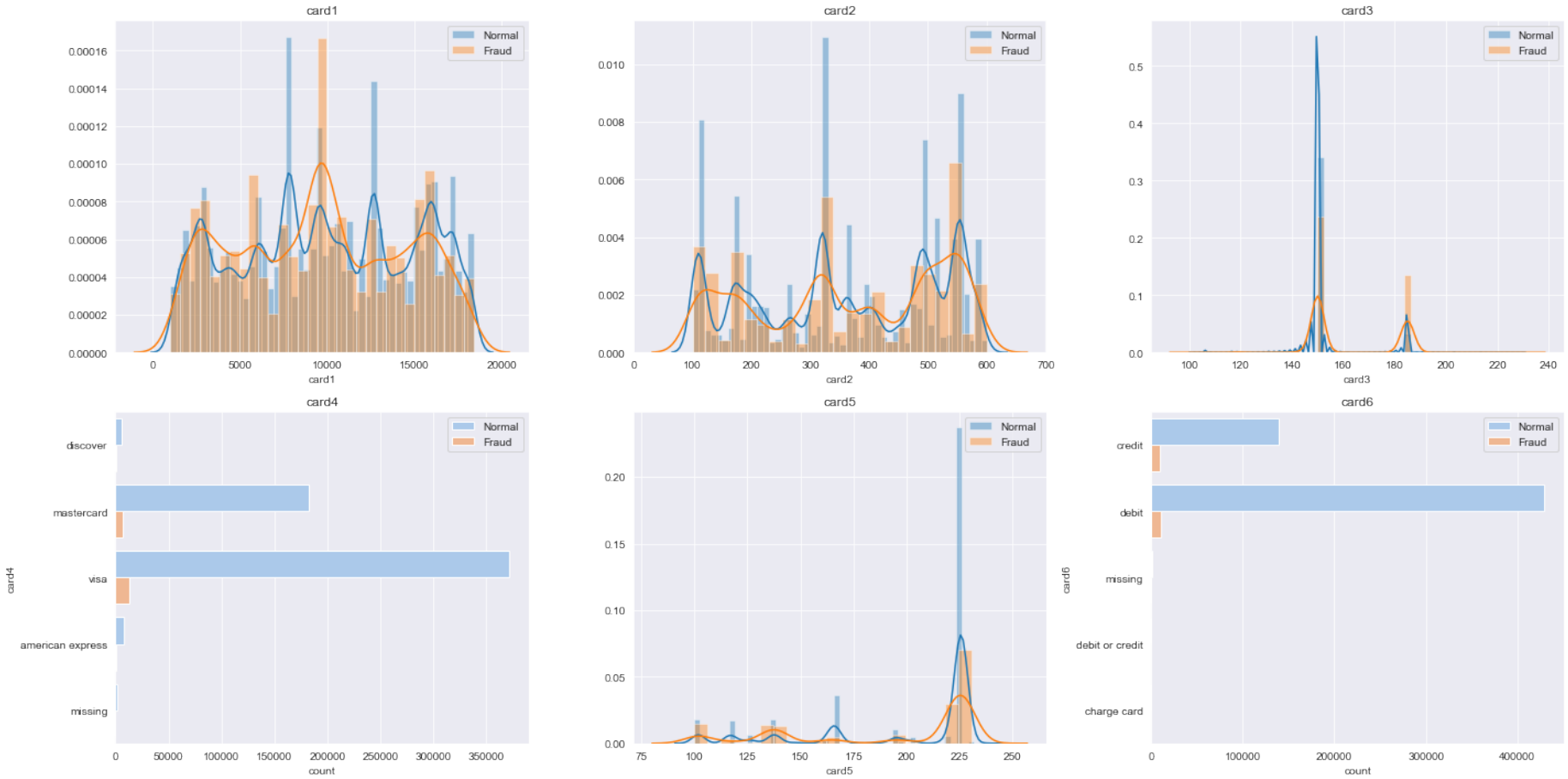




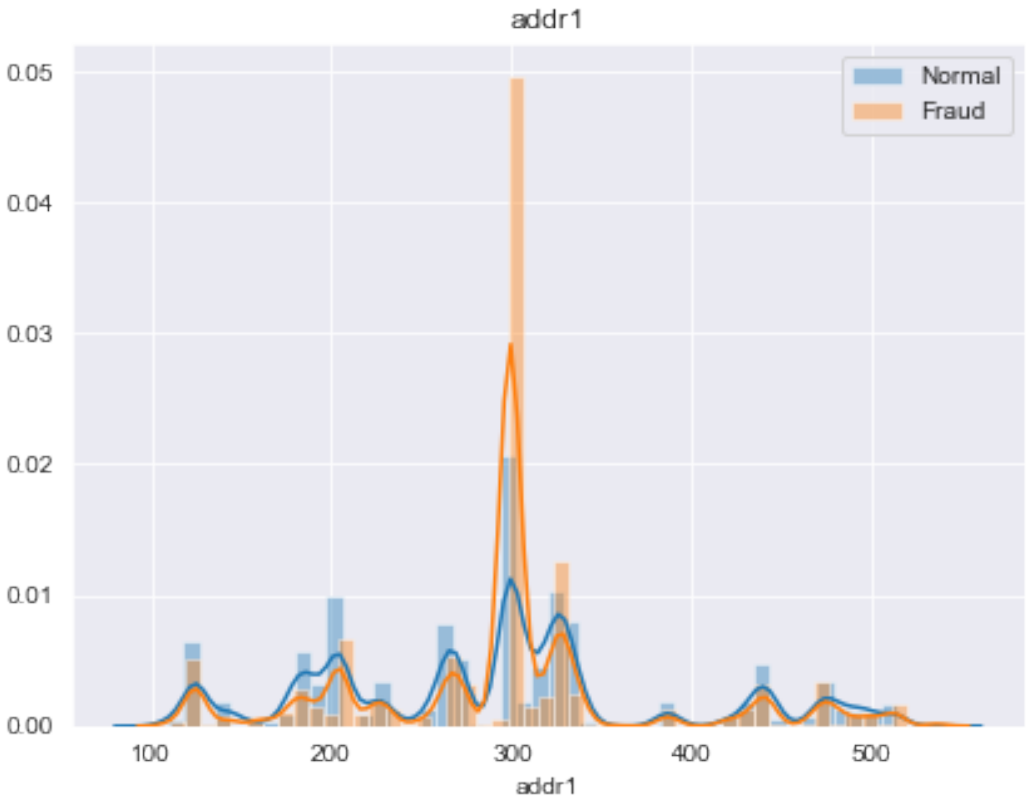
Product Codes

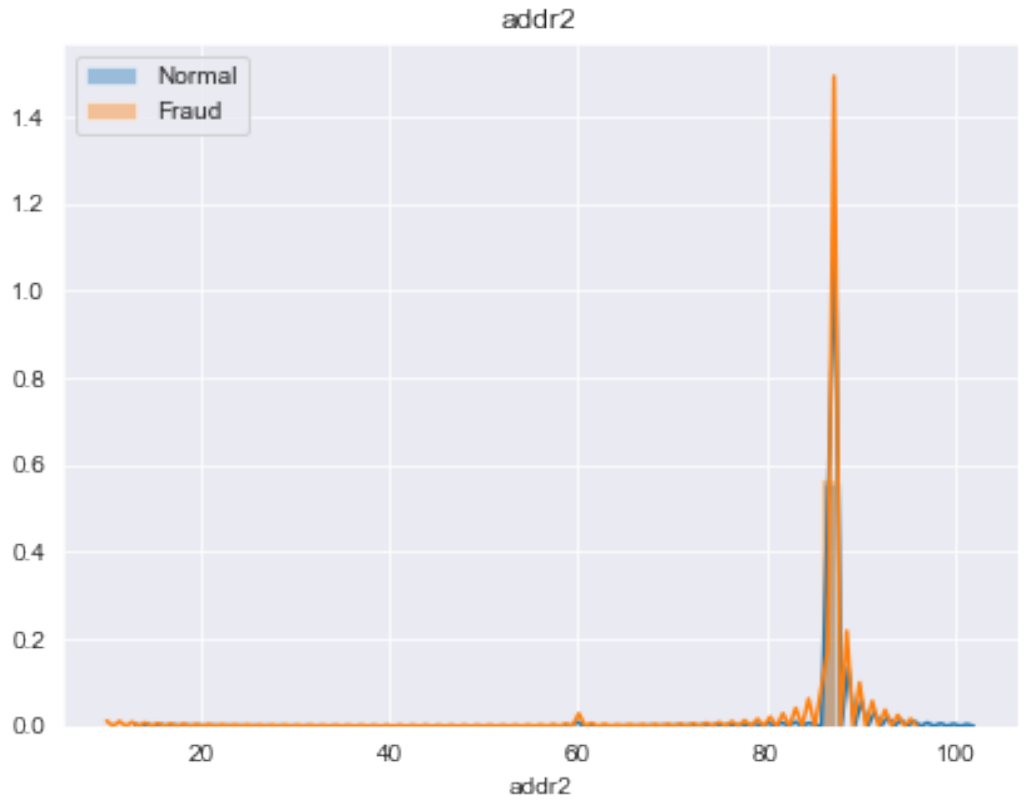


Payment Card Information

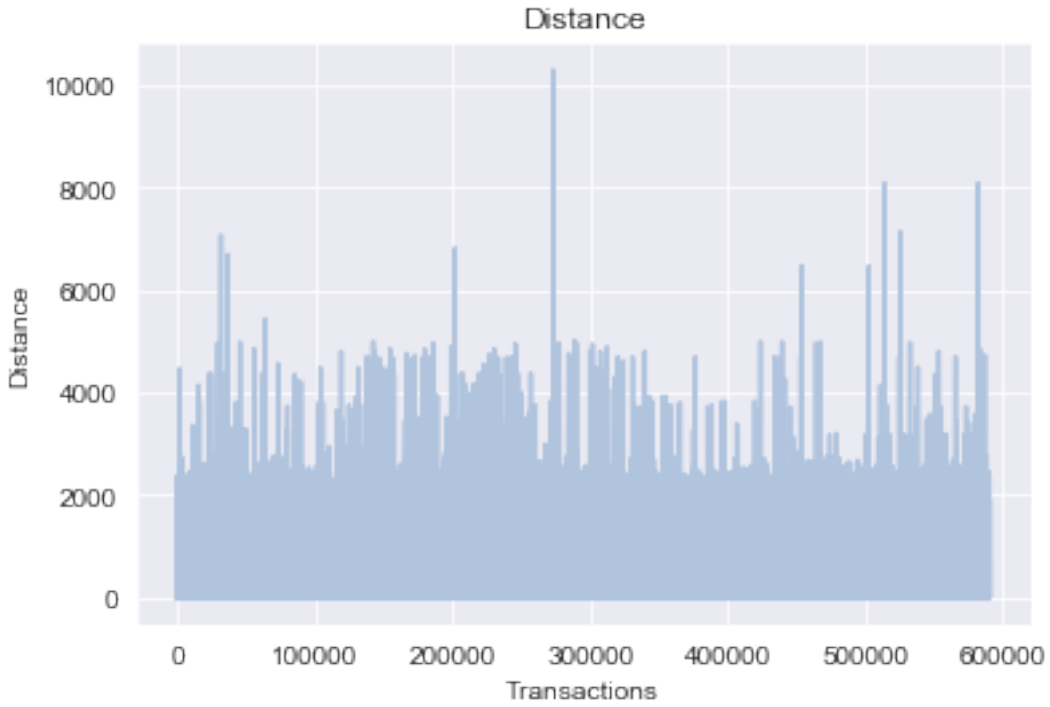


Address

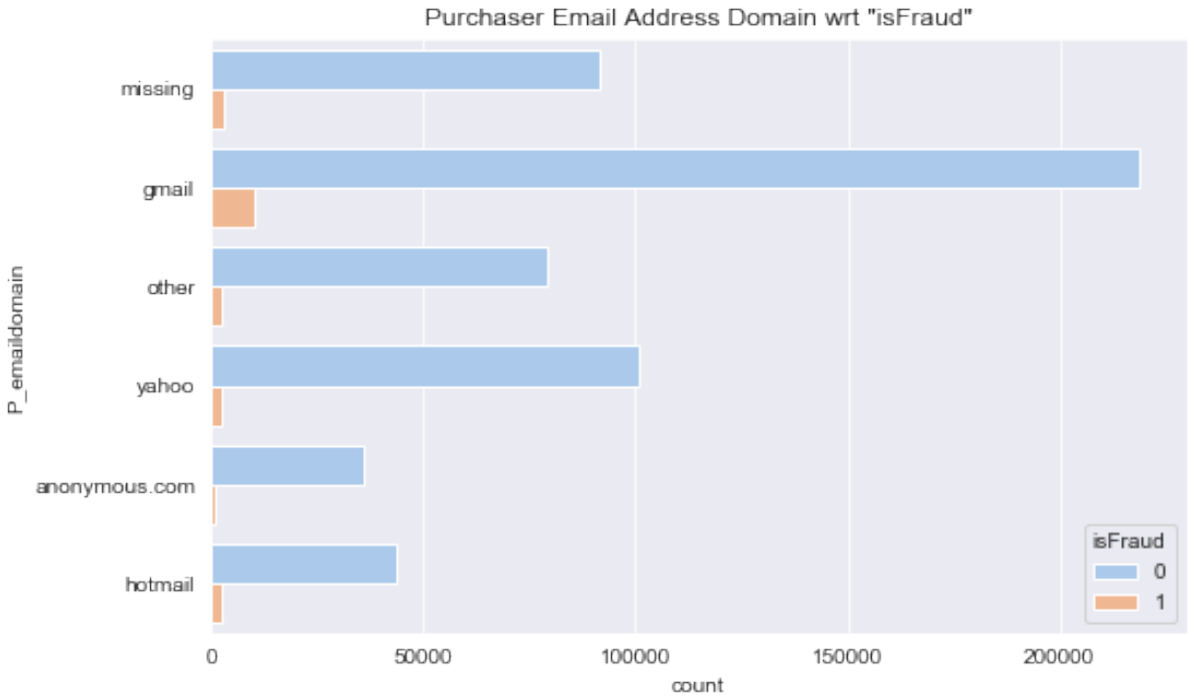

addr1
Normal
Fraud
0.05
0.04
0.03
0.02
0.01
0.00
100
200
300
400
500
addr1

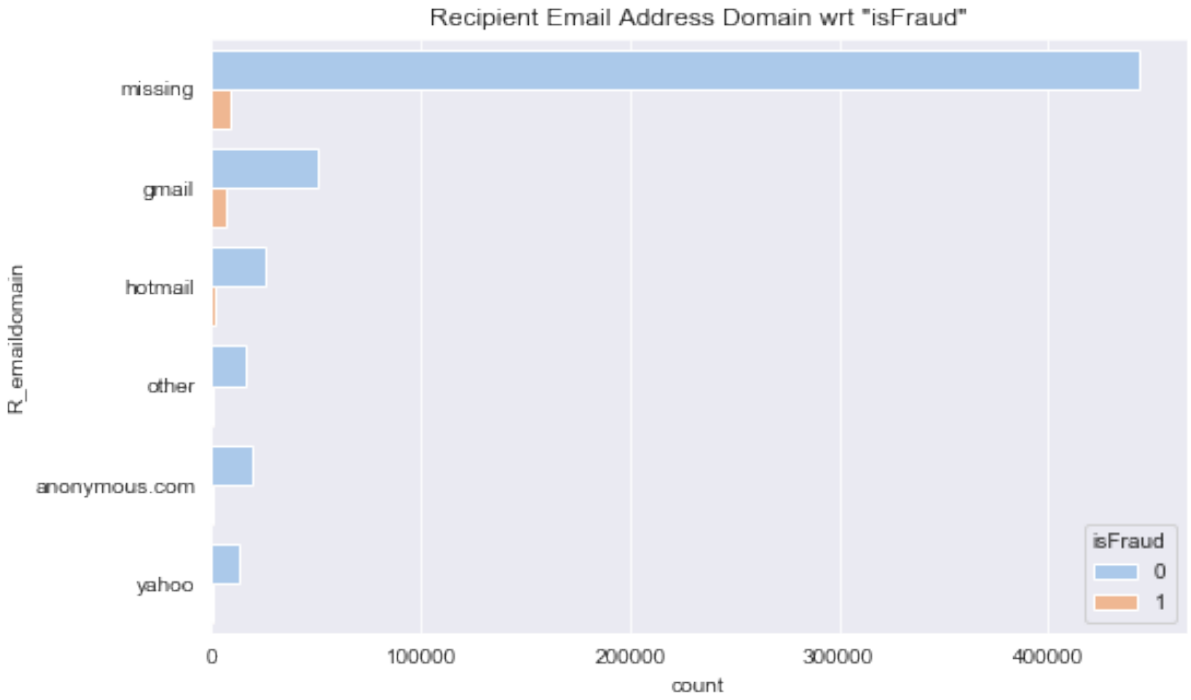

addr2
Normal
Fraud
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
100
addr2

Distance

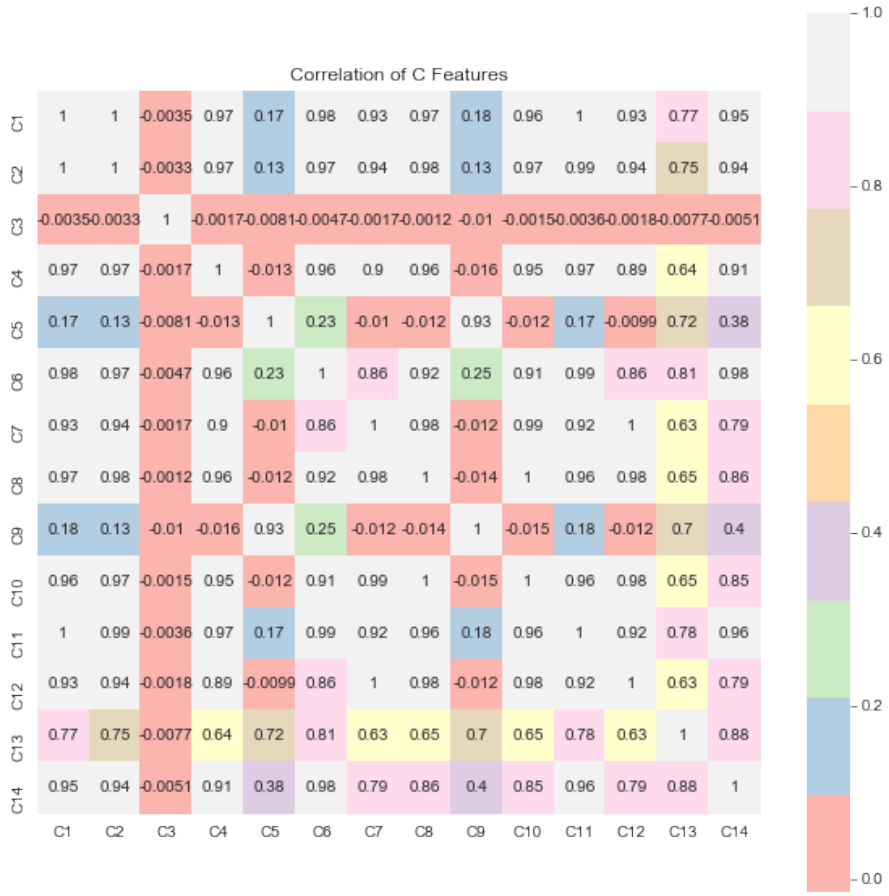

Distance
10000
8000
6000
4000
2000
0
Distance
0
100000
200000
300000
400000
500000
600000
Transactions

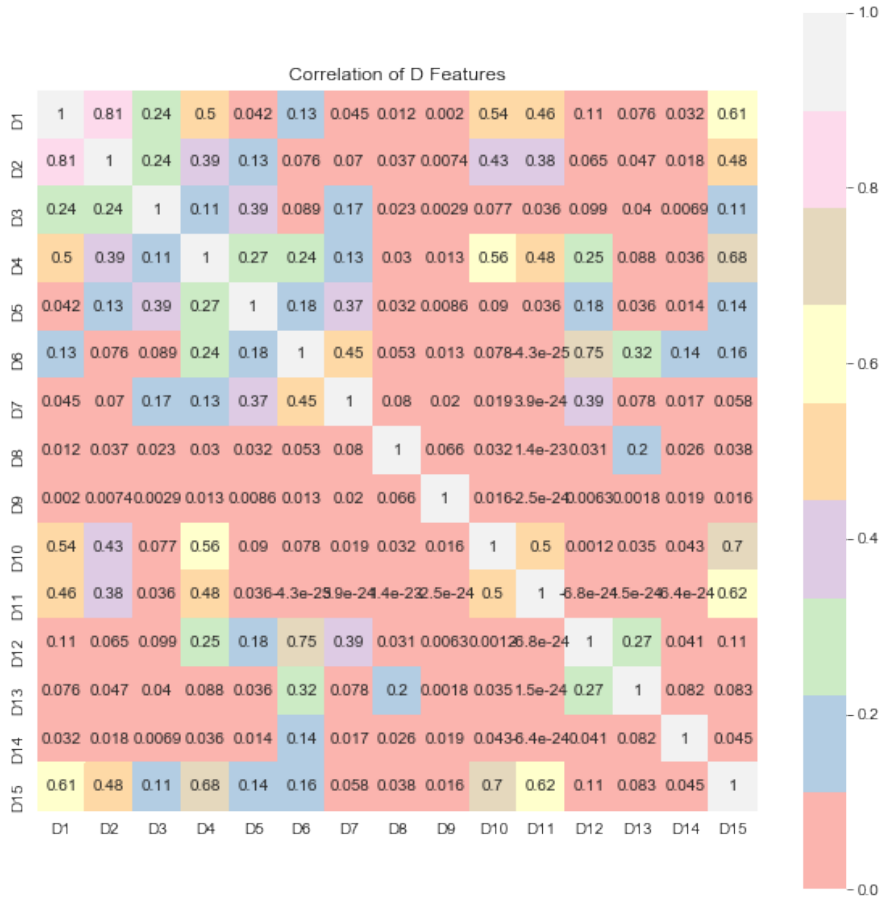
Email Domains



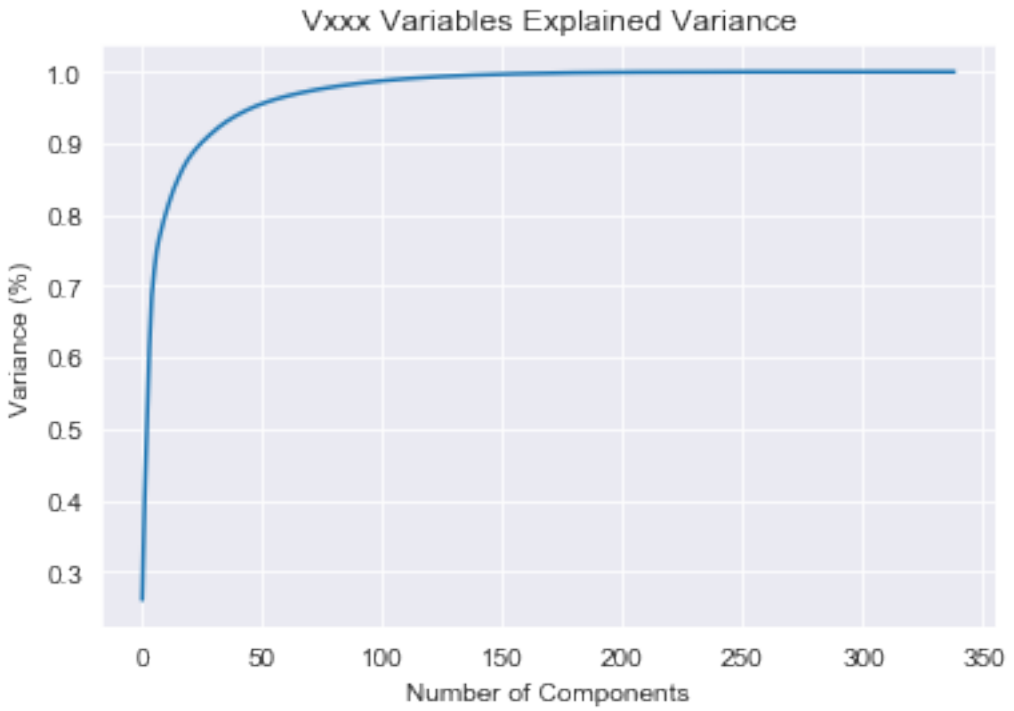


Feature Correlation



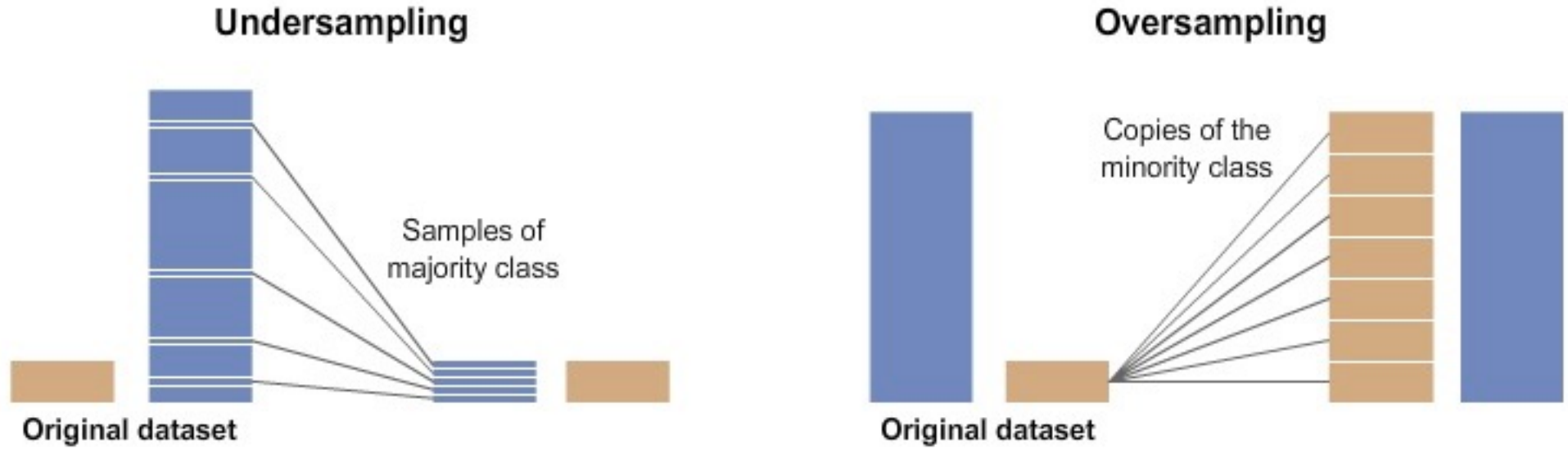


Principal Component Analysis



Class Imbalance

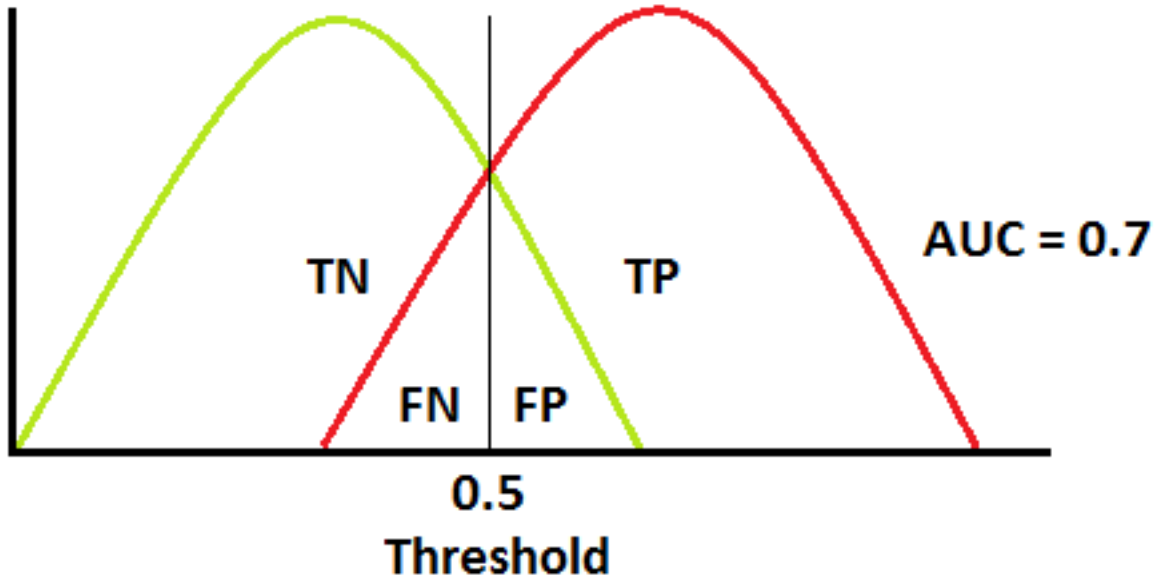
- Random Undersampling
- Synthetic Majority Oversampling Technique (SMOTE)



(source)

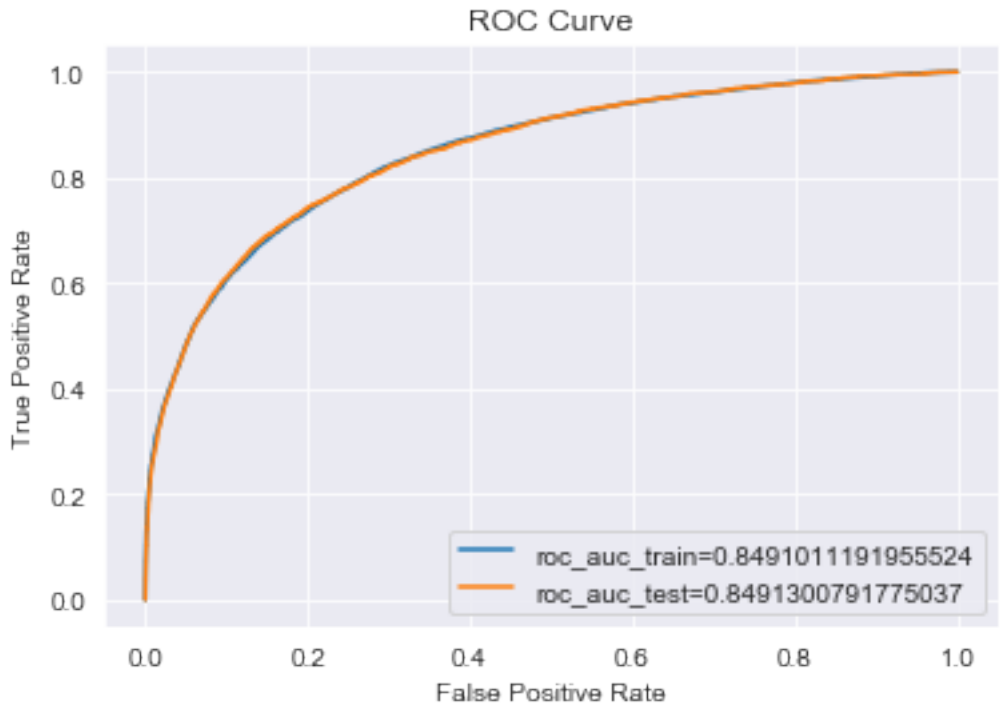
Evaluation Metrics

- Area under Receiver Operating Characteristic curve
- 0.9

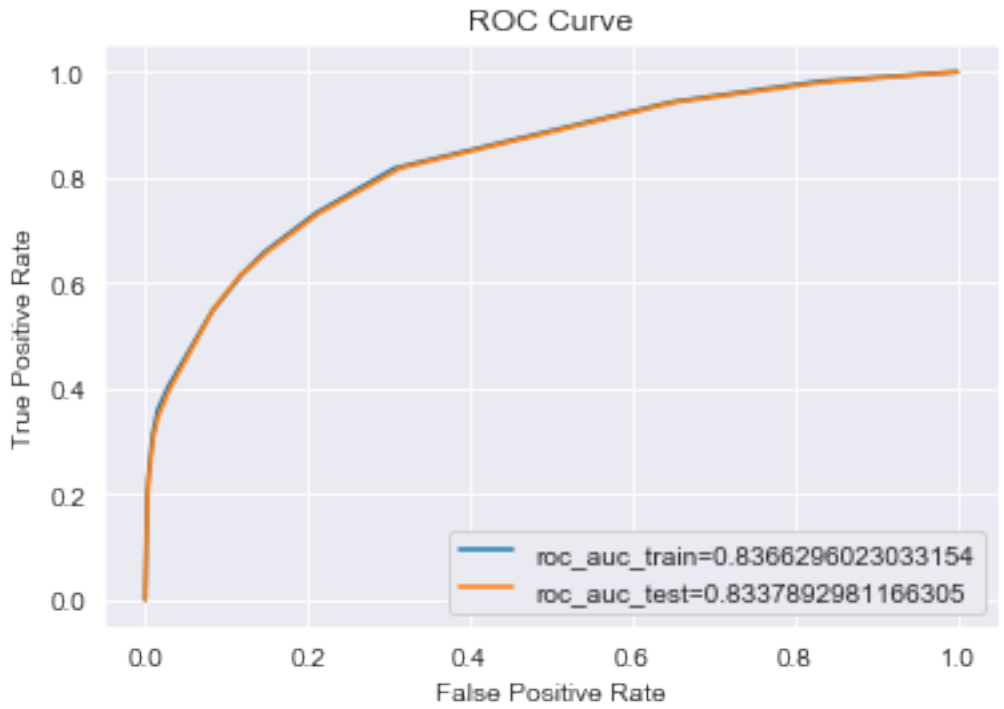


(source)

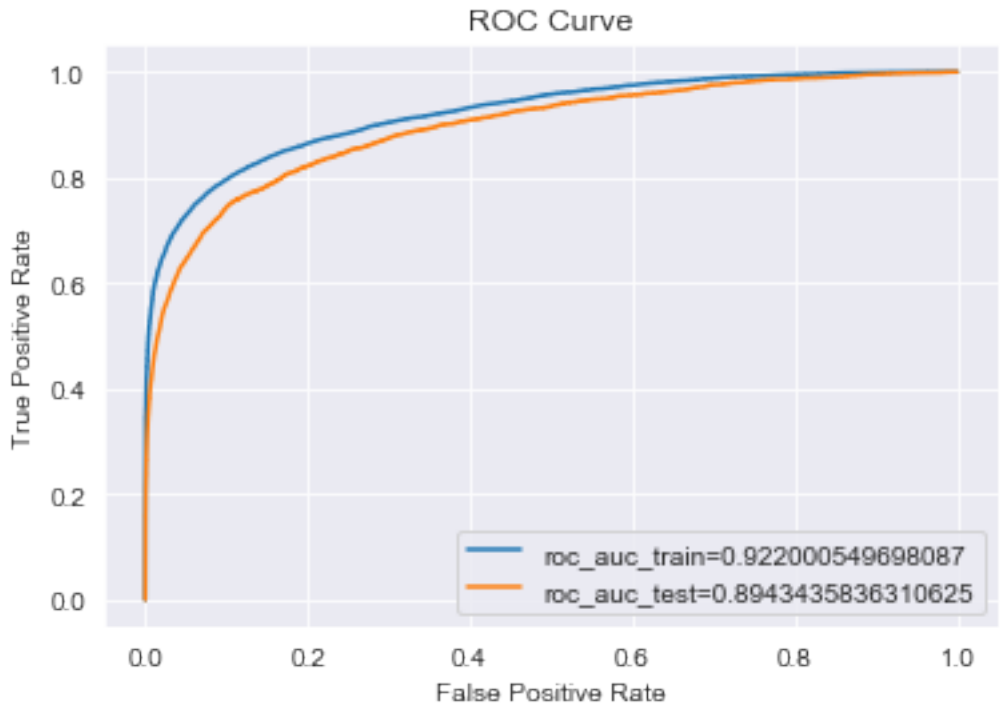
Logistic Regression



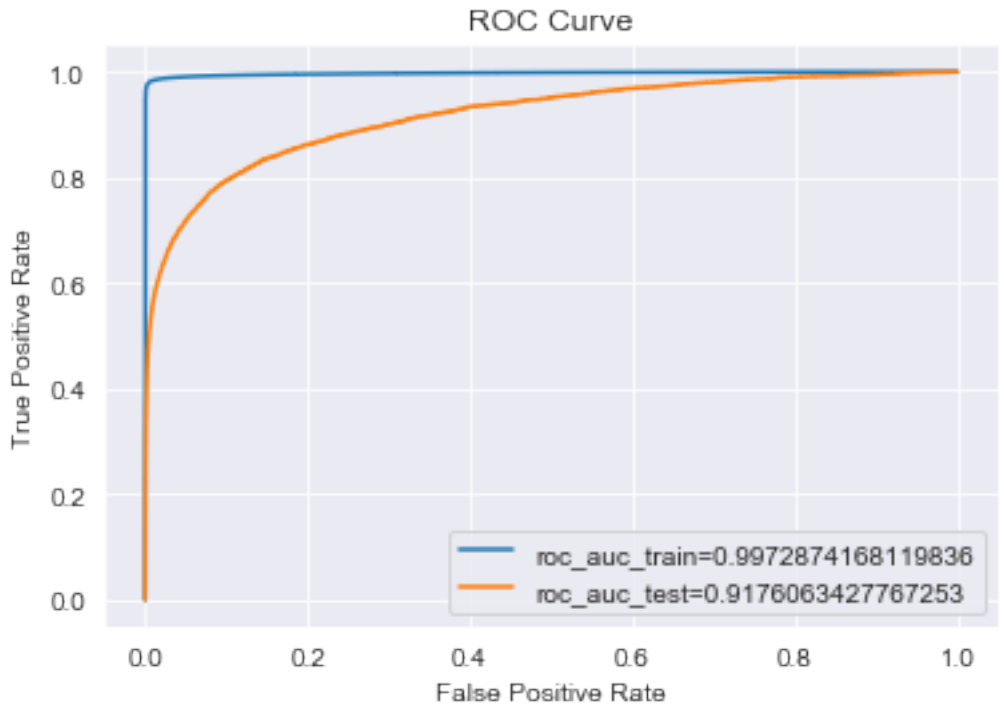
Decision Tree



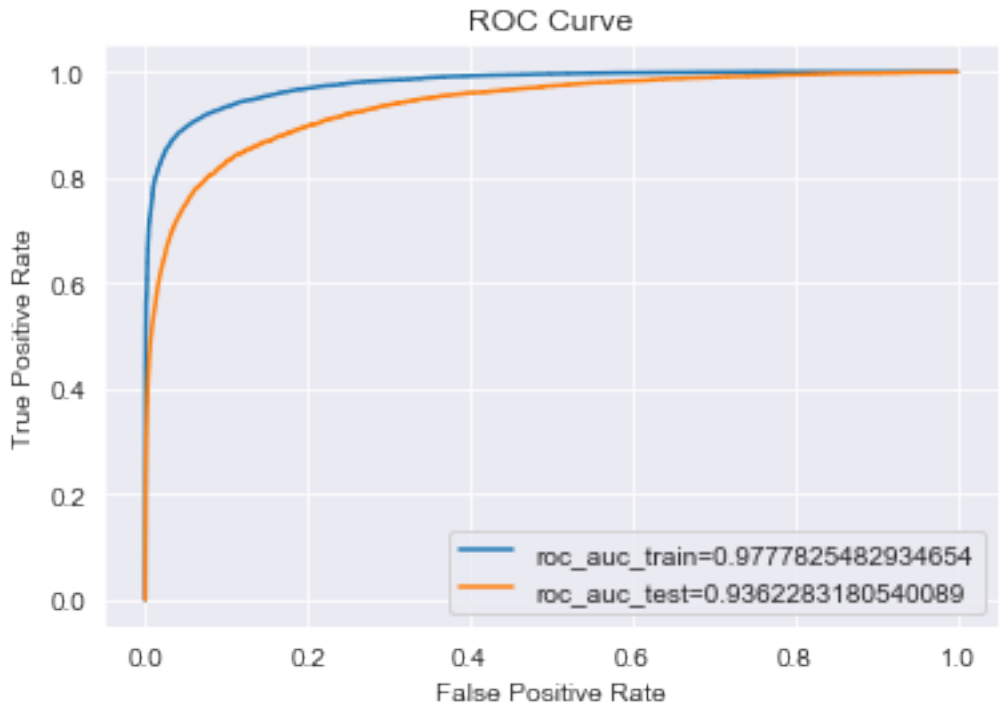
Random Forest



Gradient Boosting Classifier - SMOTE



Gradient Boosting Classifier



Important Variables (>= 0.03)

- Vesta engineered features (actual meaning is masked):
  - 1, PCA (e.g. ranking, counting, other entity relations)
  - C features (counting, e.g. how many address found)
  - D3 (time delta, such as days between previous transaction)
- Transaction Amount

Summary

ROC AUC	Precision	Recall	Type I Error	Type II Error
0.9362	0.21	0.85	0.112	0.005

Q&A