

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

Data Overview

- Two datasets: transaction and identity data
- 590540 observations, 434 variables

Workflow

I. Data Preparation:

- Data Cleaning
- Exploratory Data Analysis
- Feature Engineering

II. Modeling:

- Class Imbalance
- Model Optimization
- Model Evaluation

Data Preparation

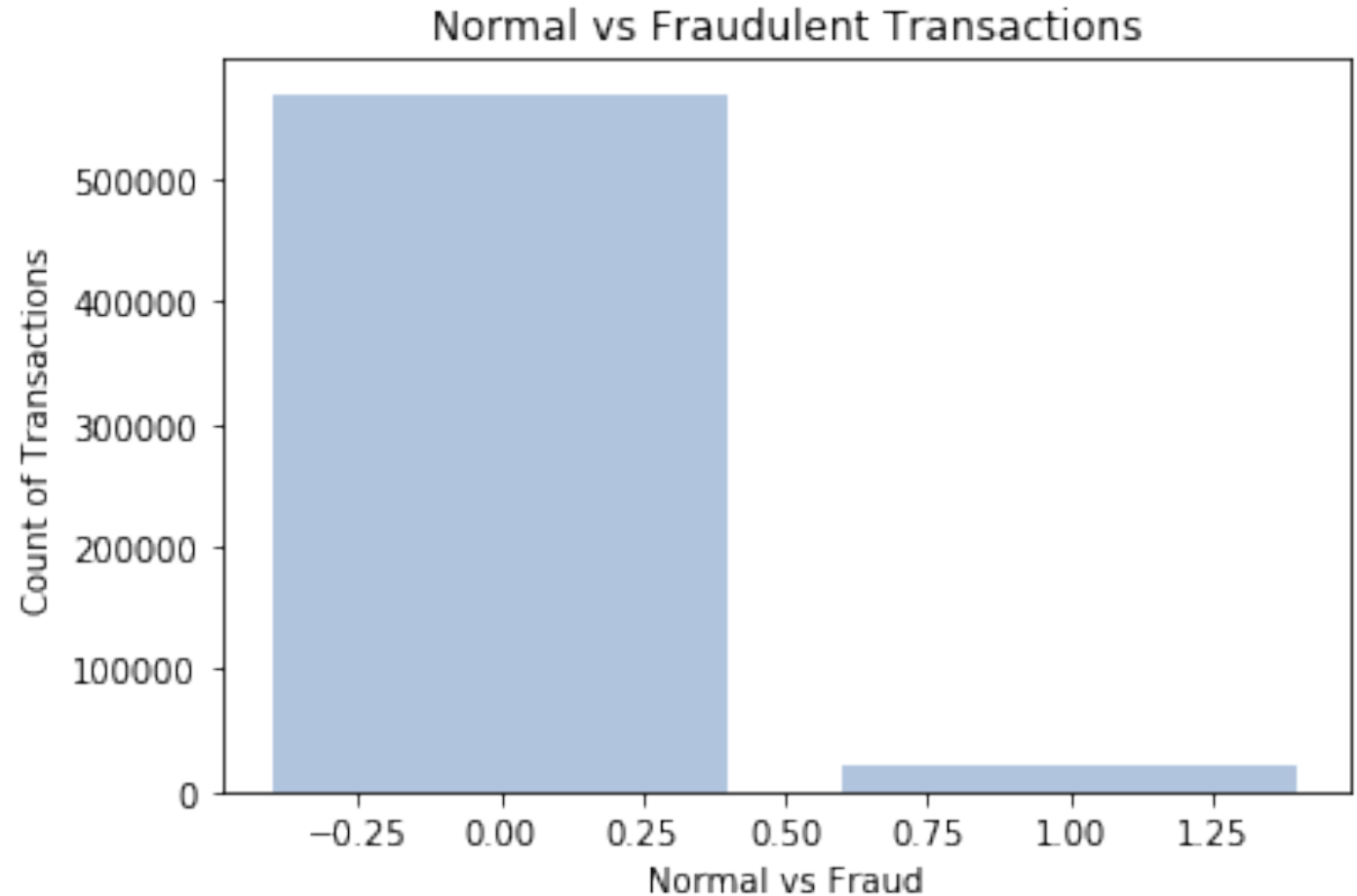
- Address missing data
- Limit values of outliers
- Transform variables
- Reduce dimensionality

Data Preparation

- Continuous variables:
 - Transaction time deltas
 - Transaction amount
 - Distance
 - Vesta engineered features
- Categorical variables:
 - Product codes
 - Payment card information
 - Address
 - Email domains
 - Identity information

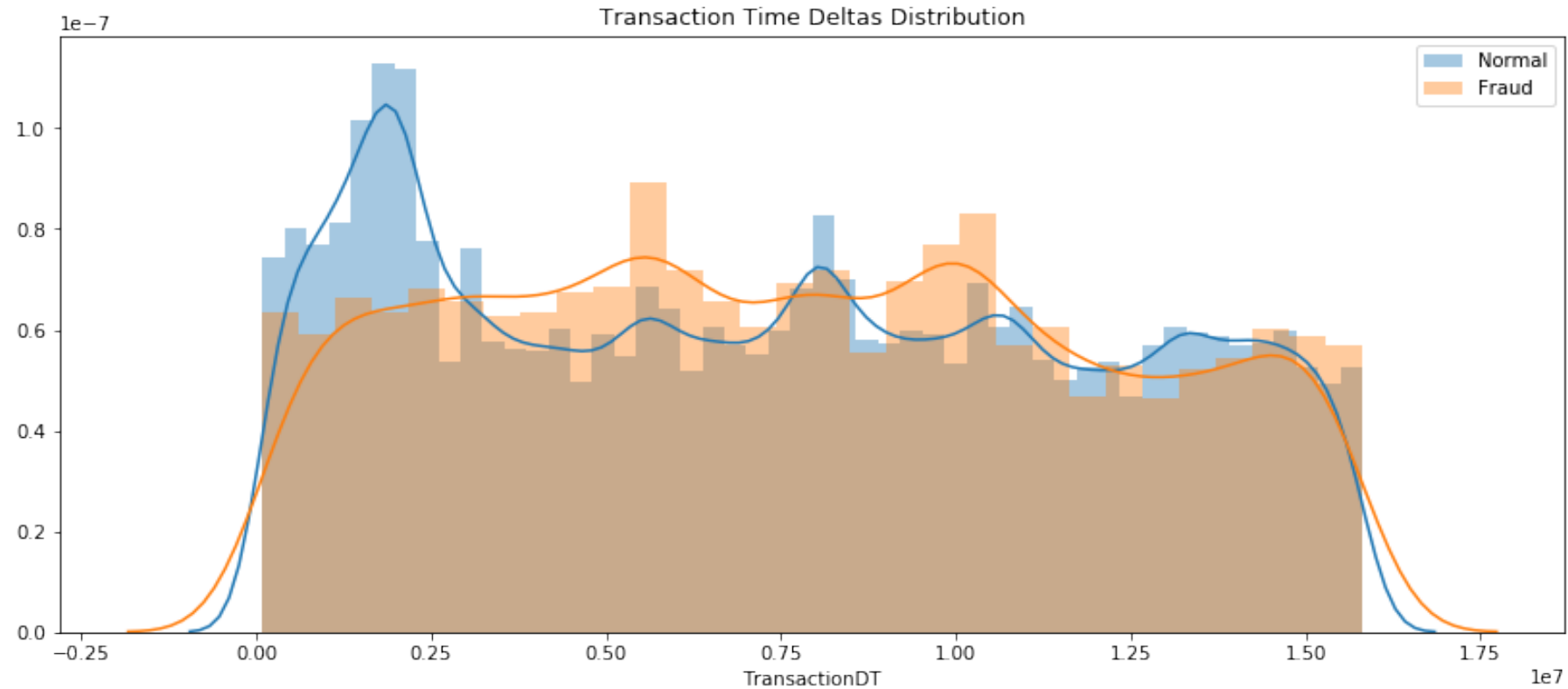
Data Preparation

- Target: isFraud
- 3.6%
- Class Imbalance



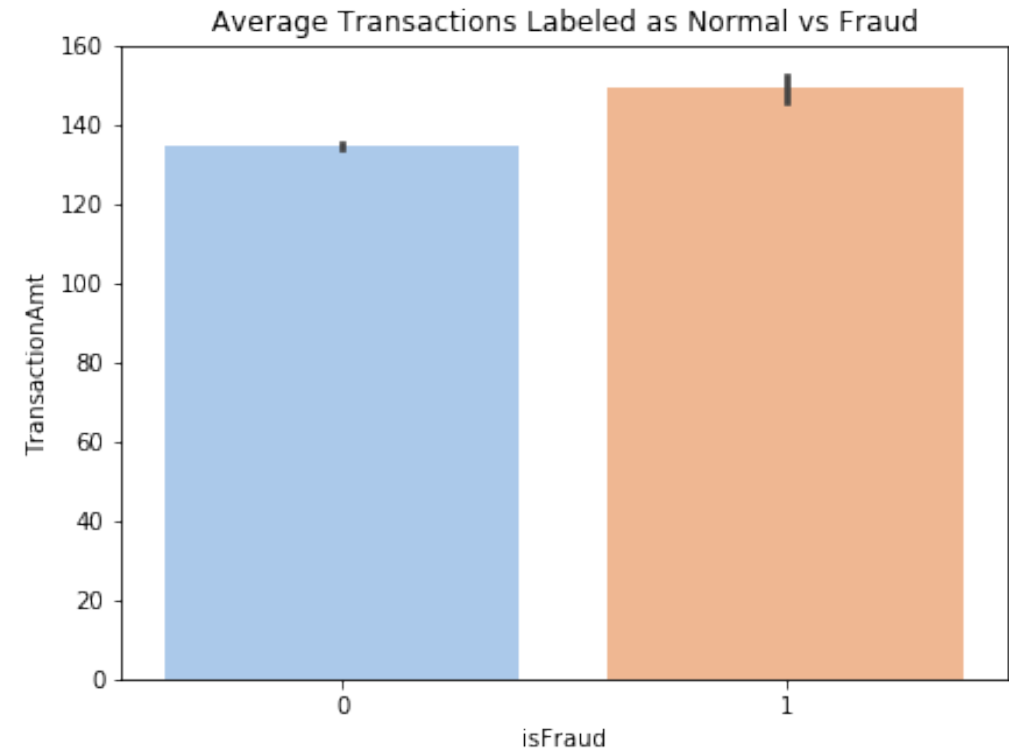
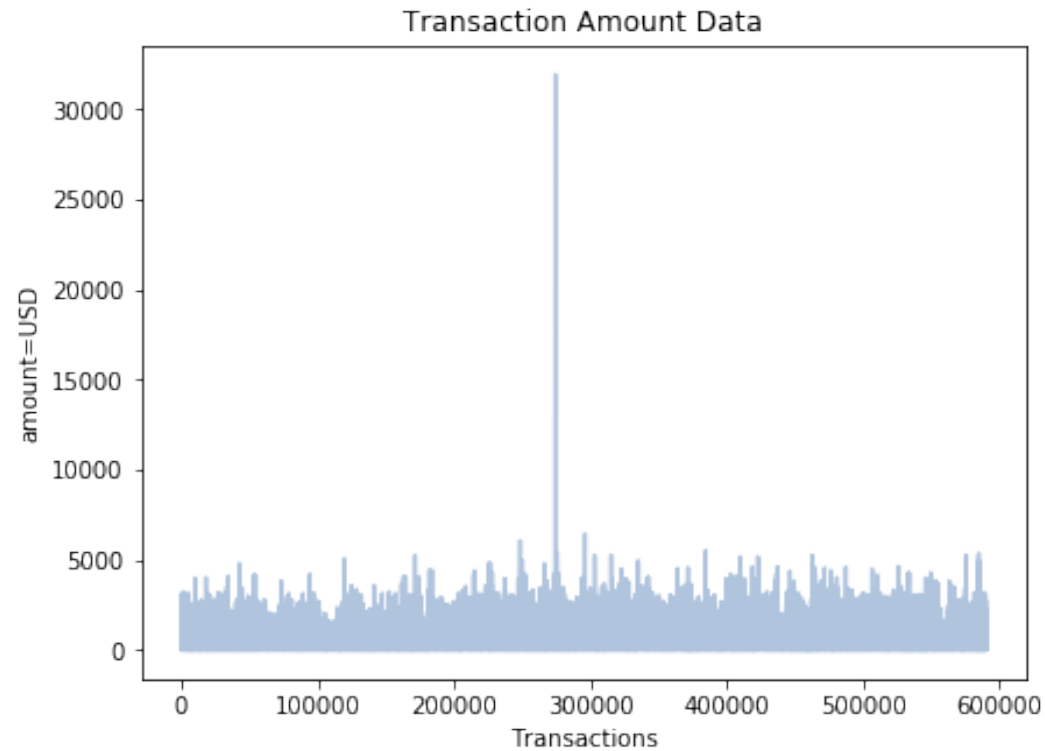
Data Preparation

Transaction Time Deltas



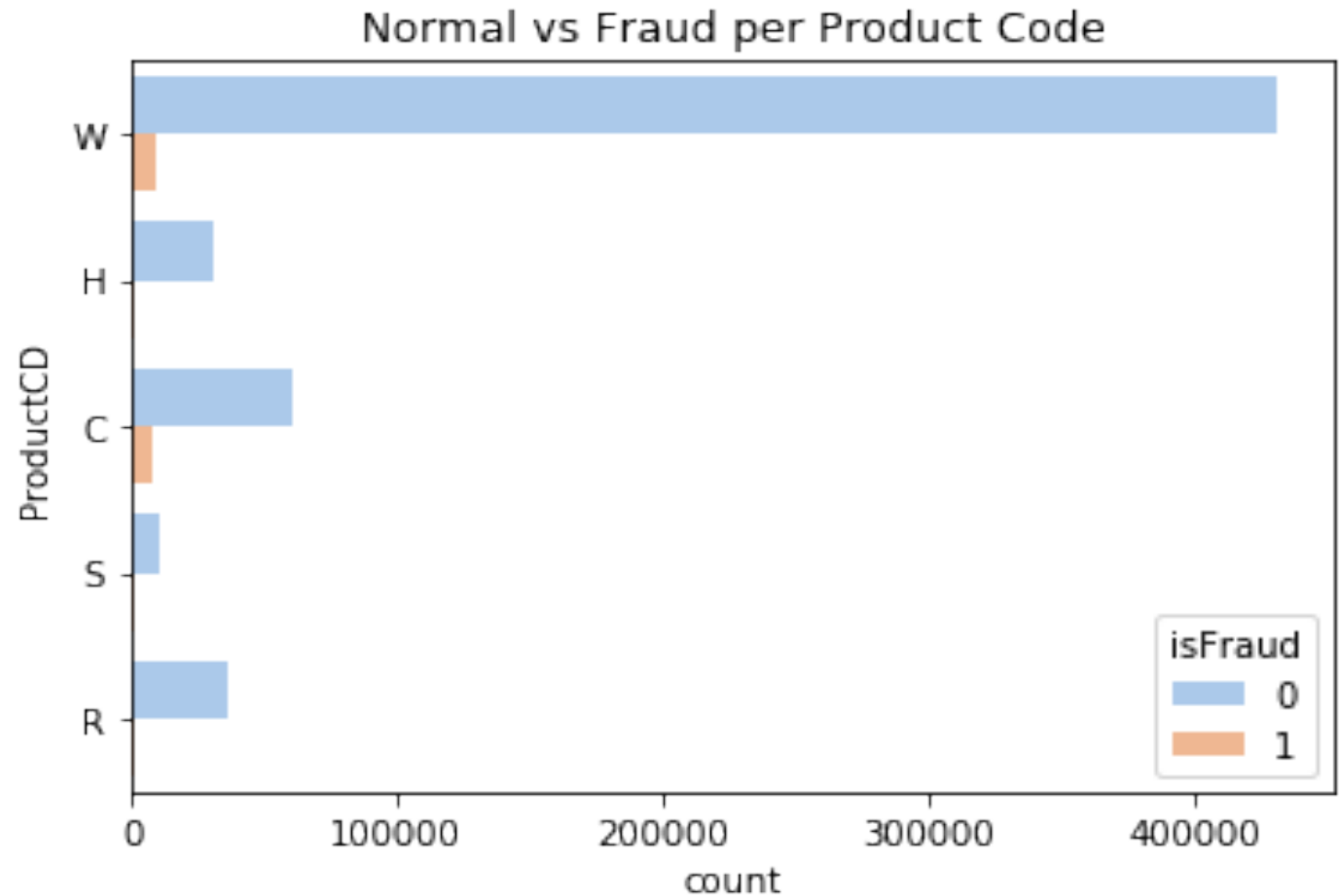
Data Preparation

Transaction Amount



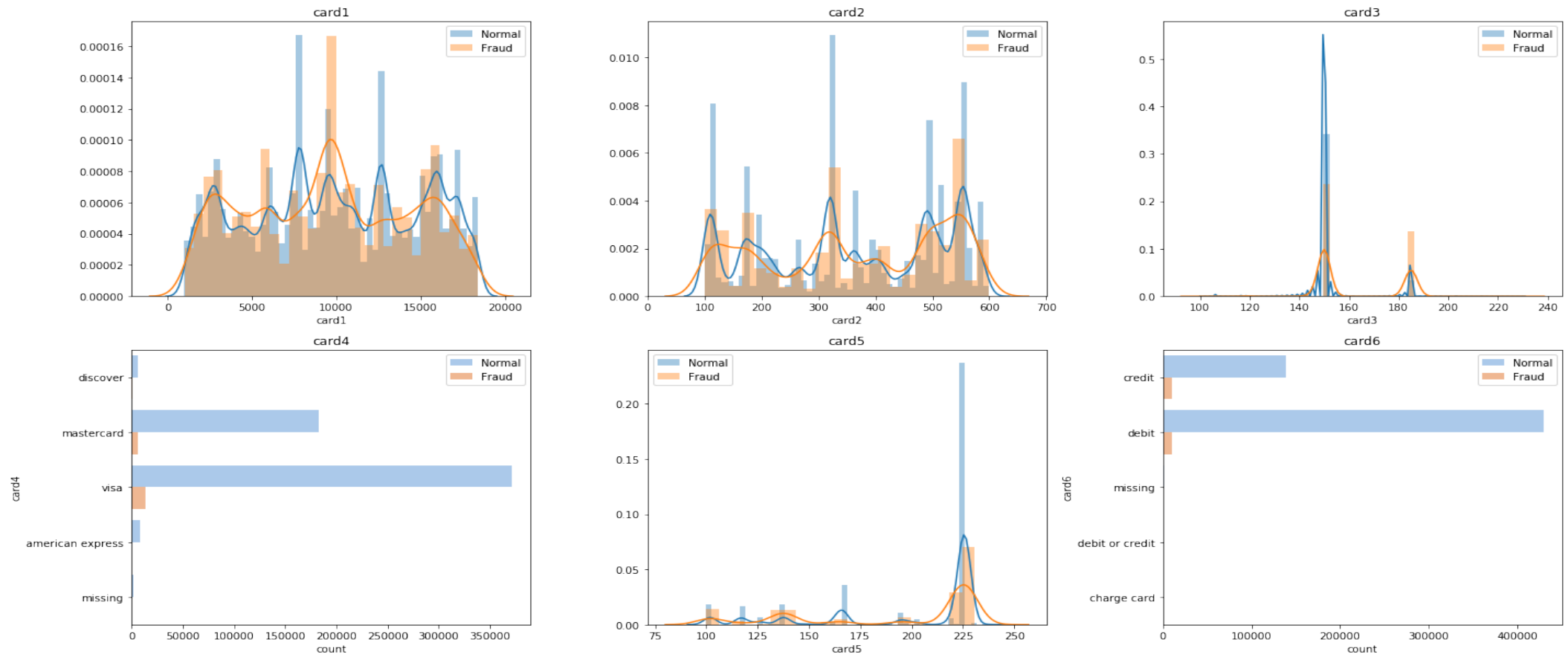
Data Preparation: Product Code

Product Codes



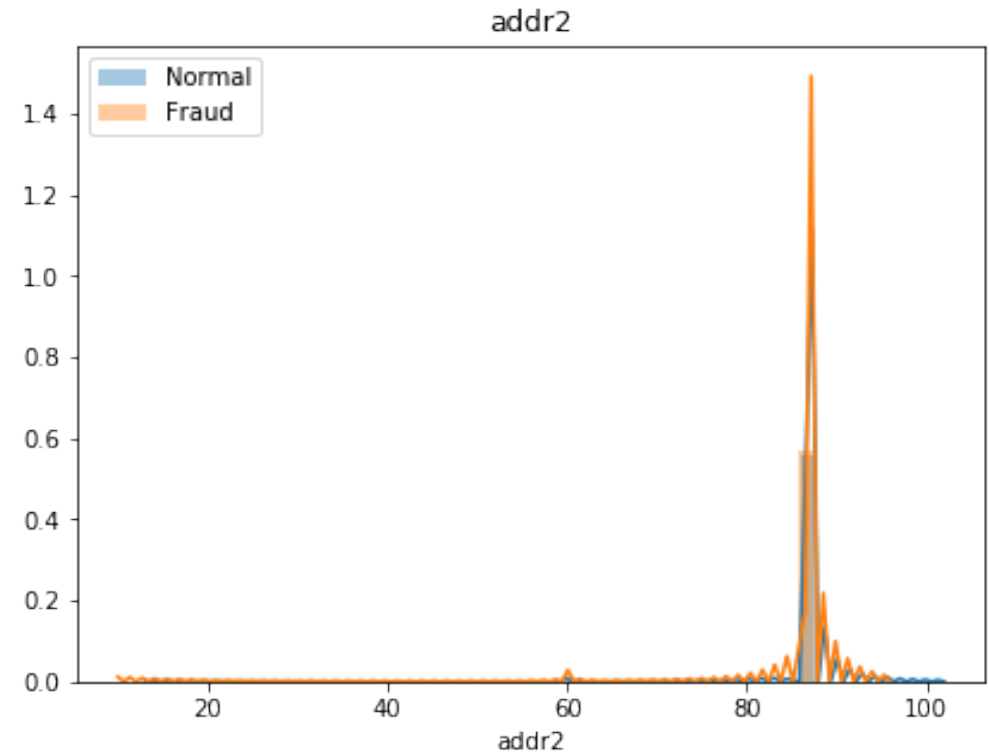
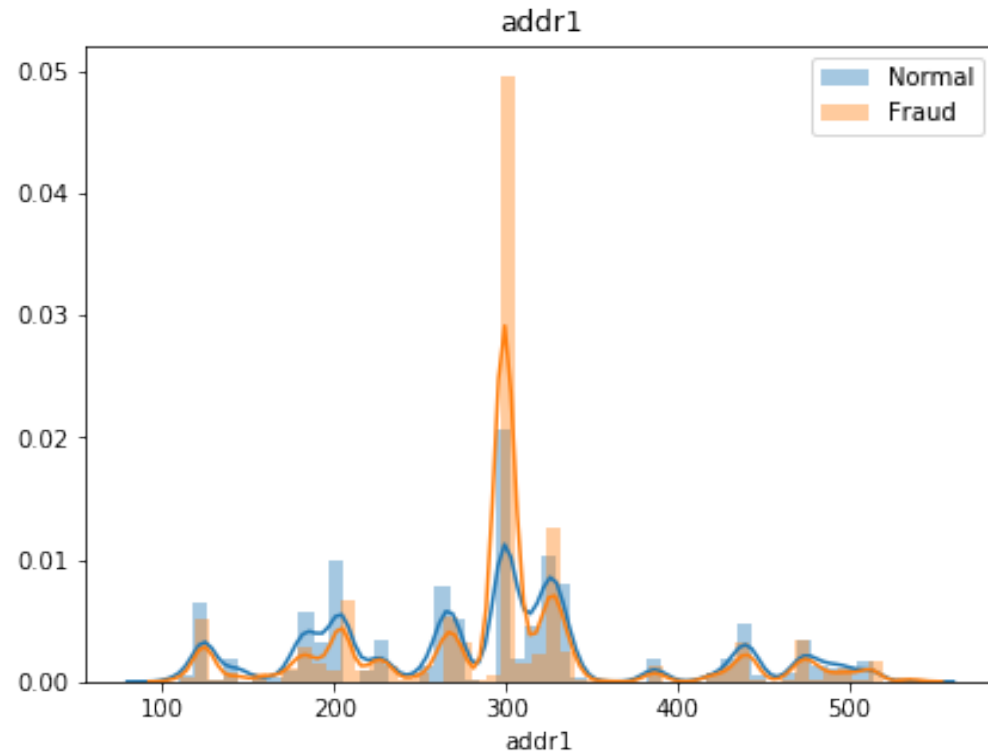
Data Preparation

Payment Card Information



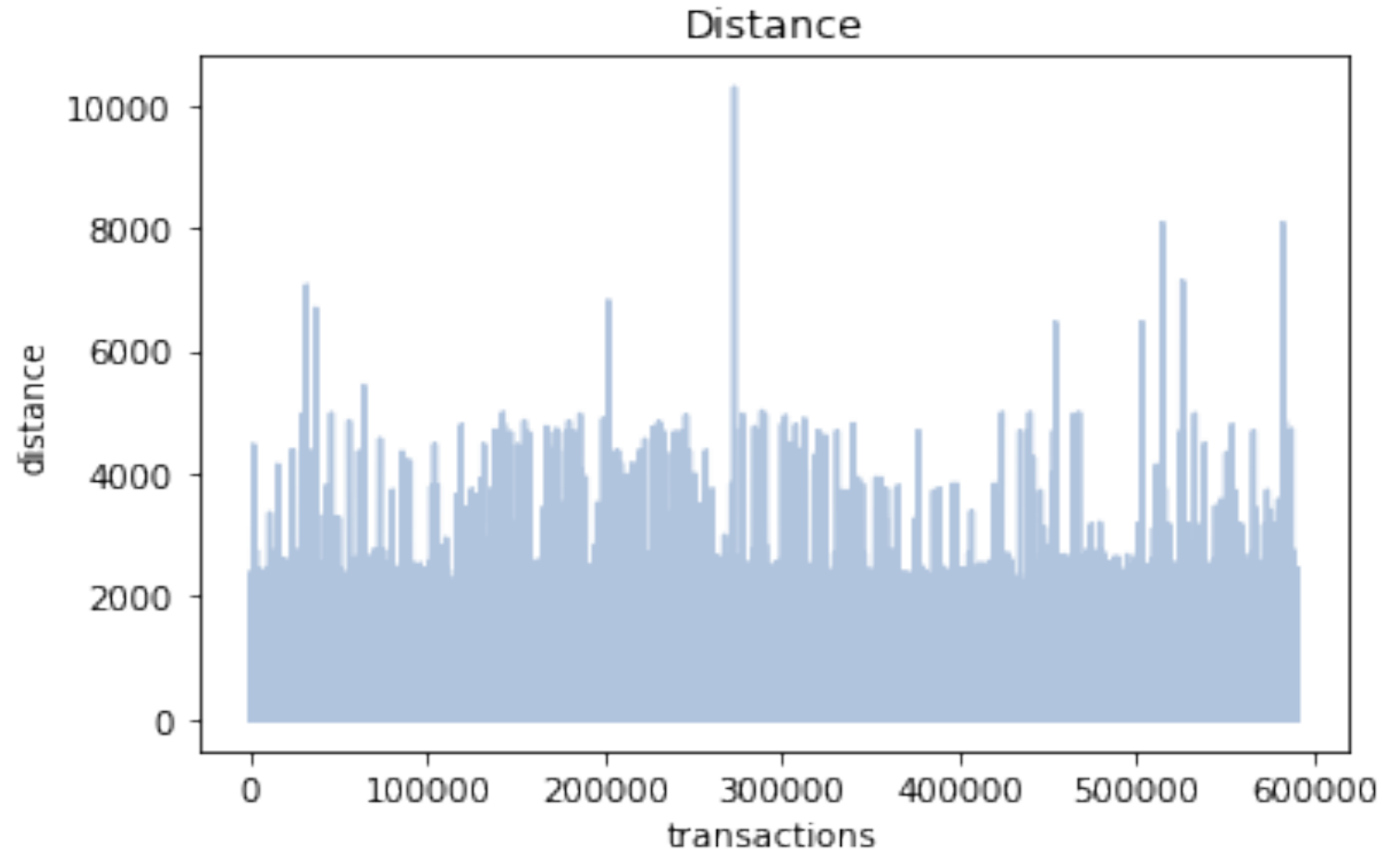
Data Preparation

Address



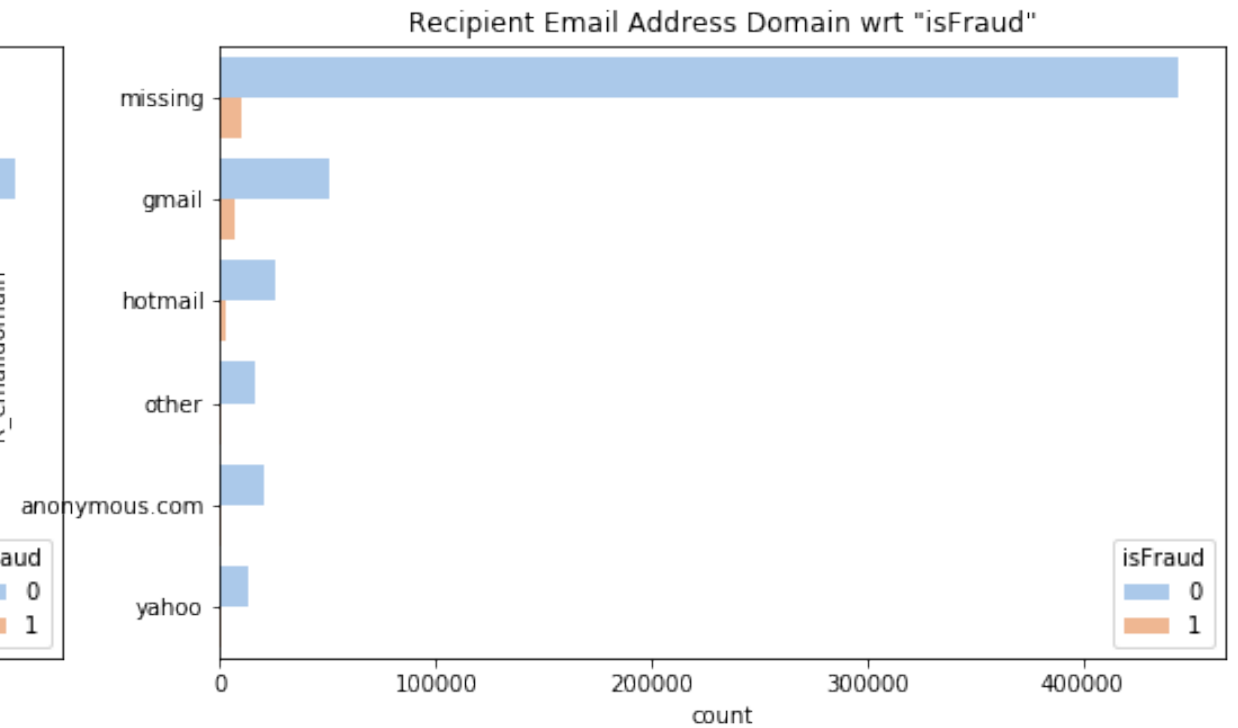
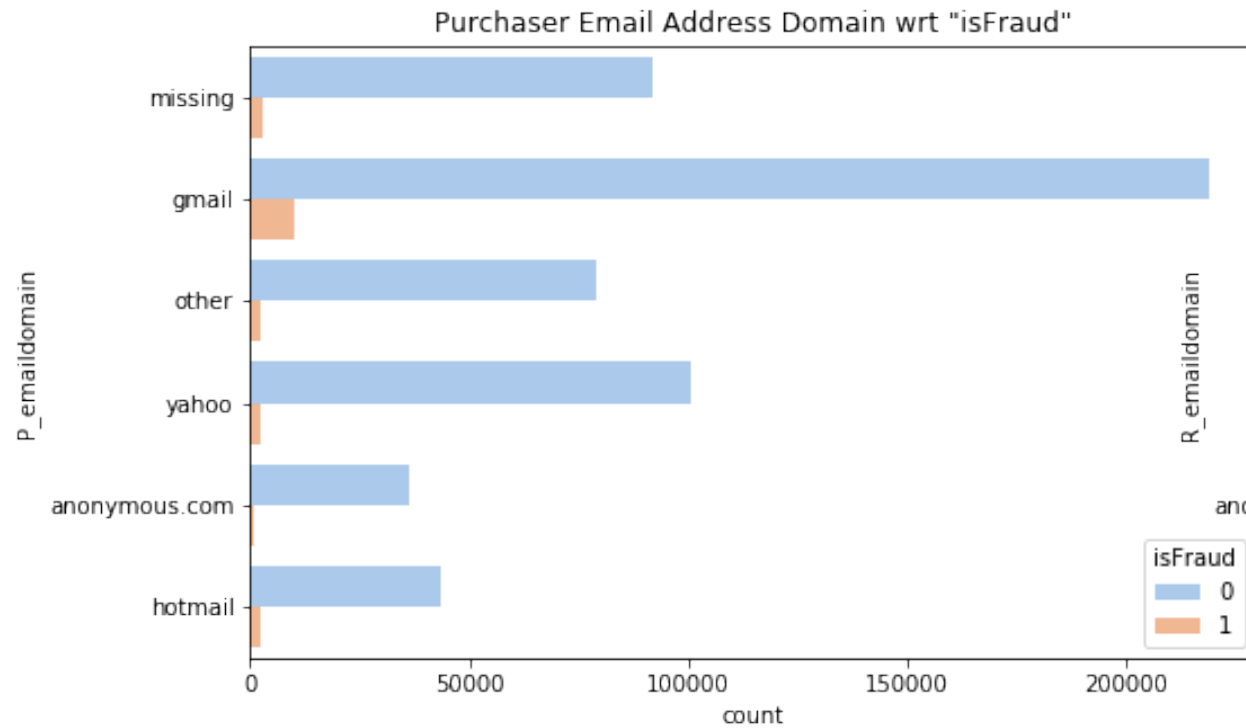
Data Preparation

Distance



Data Preparation

Email domains



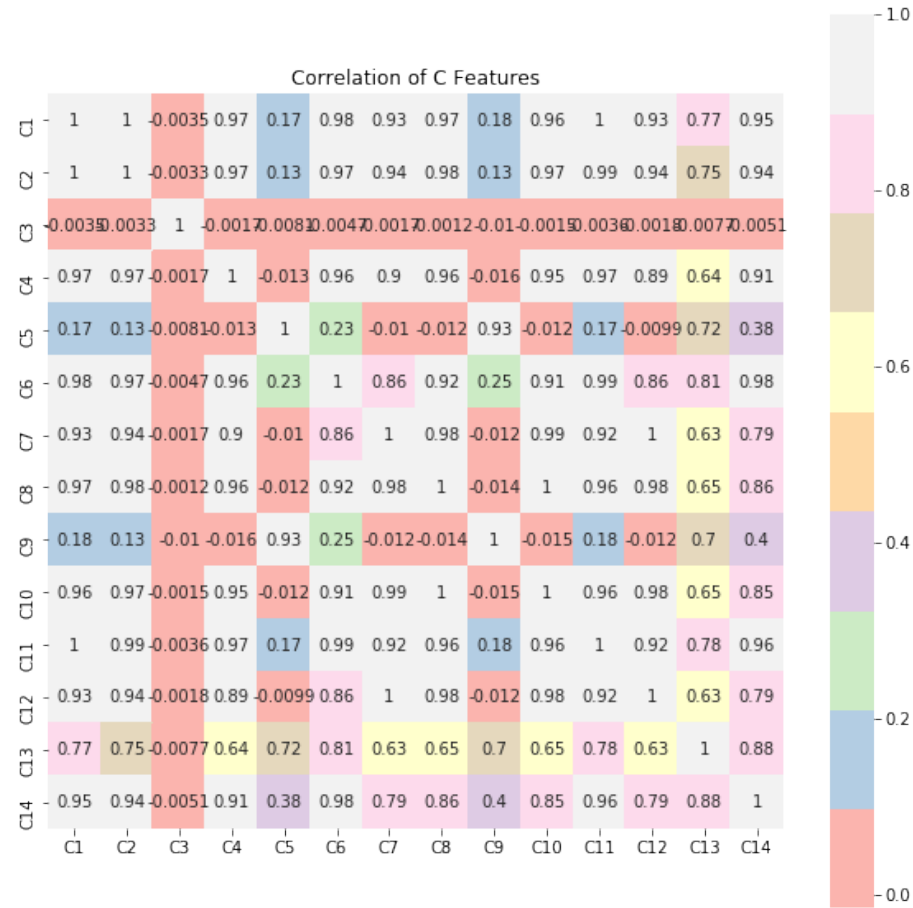
Data Preparation

Vesta-engineered features:

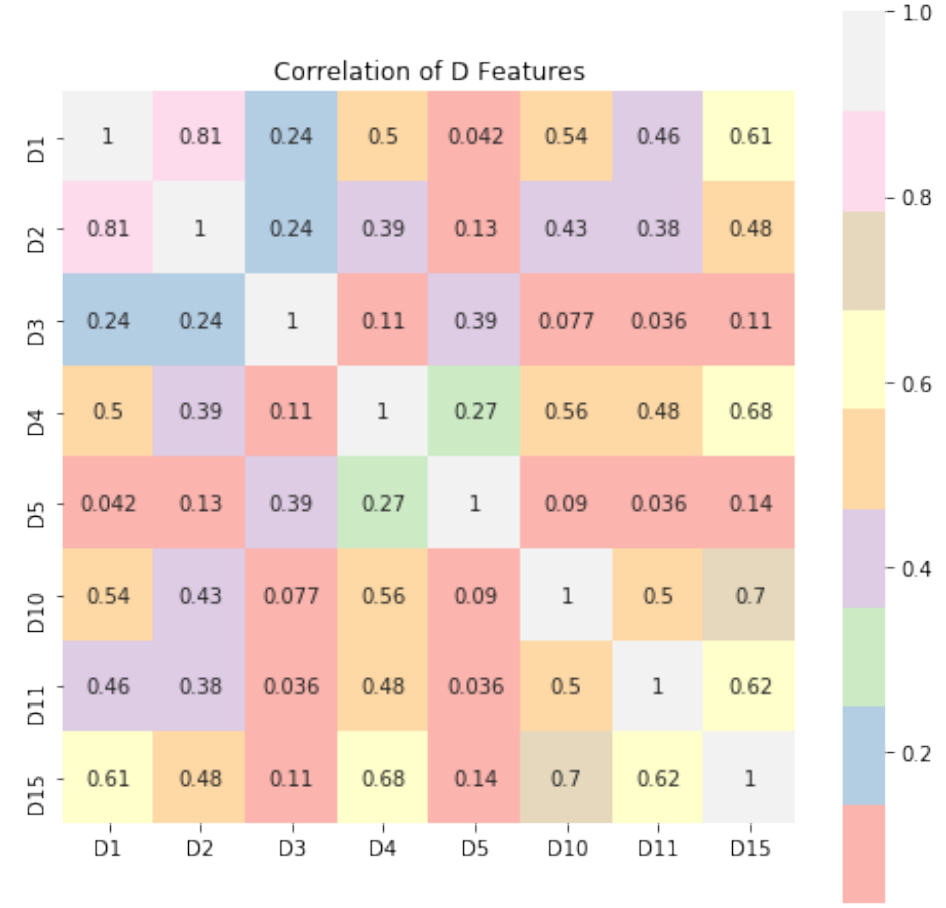
- C variables
- D variables
- M variables
- V variables

Data Preparation

Correlation of C Features



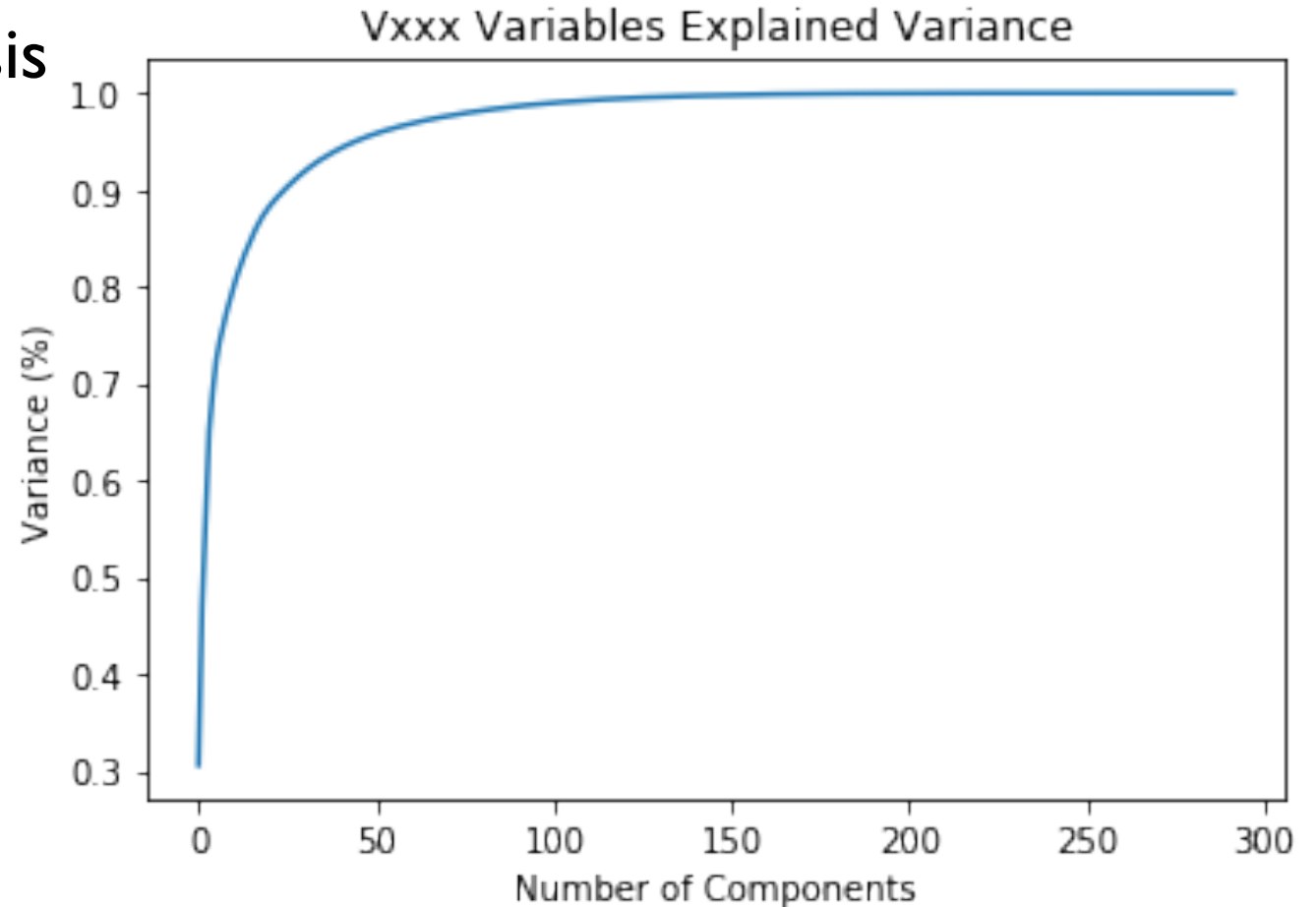
Correlation of D Features



Data Preparation

Principal Component Analysis

- 292 variables
- reduced to 25 features



Data Preparation

Identity/Device Information:

- Over 75% of missing values

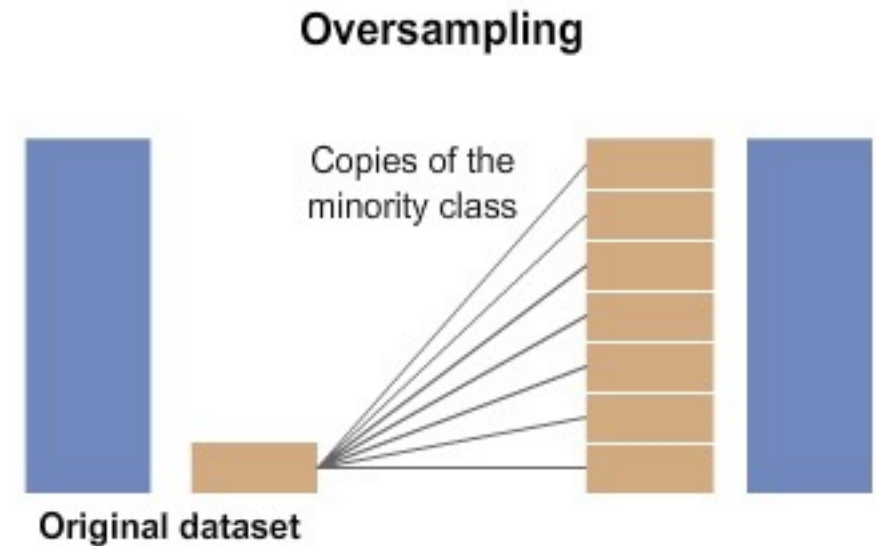
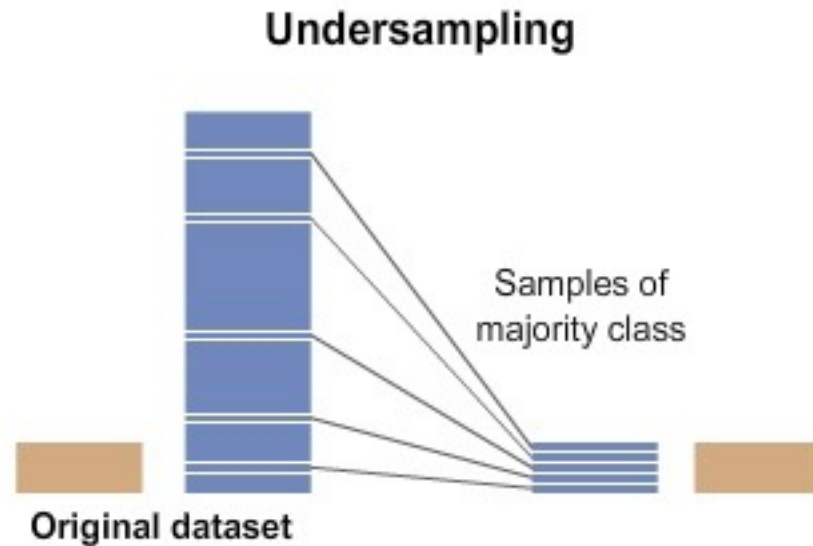
Data Preparation Summary

- Data is numeric
- 91 features
- Some features are highly correlated
- Some variables had a lot of missing values

Modeling

Class Imbalance:

- Random Undersampling
- Synthetic Majority Oversampling Technique (SMOTE)



Modeling

Evaluation Metrics:

- Area under Receiver Operating Characteristic curve
- Precision
- Recall

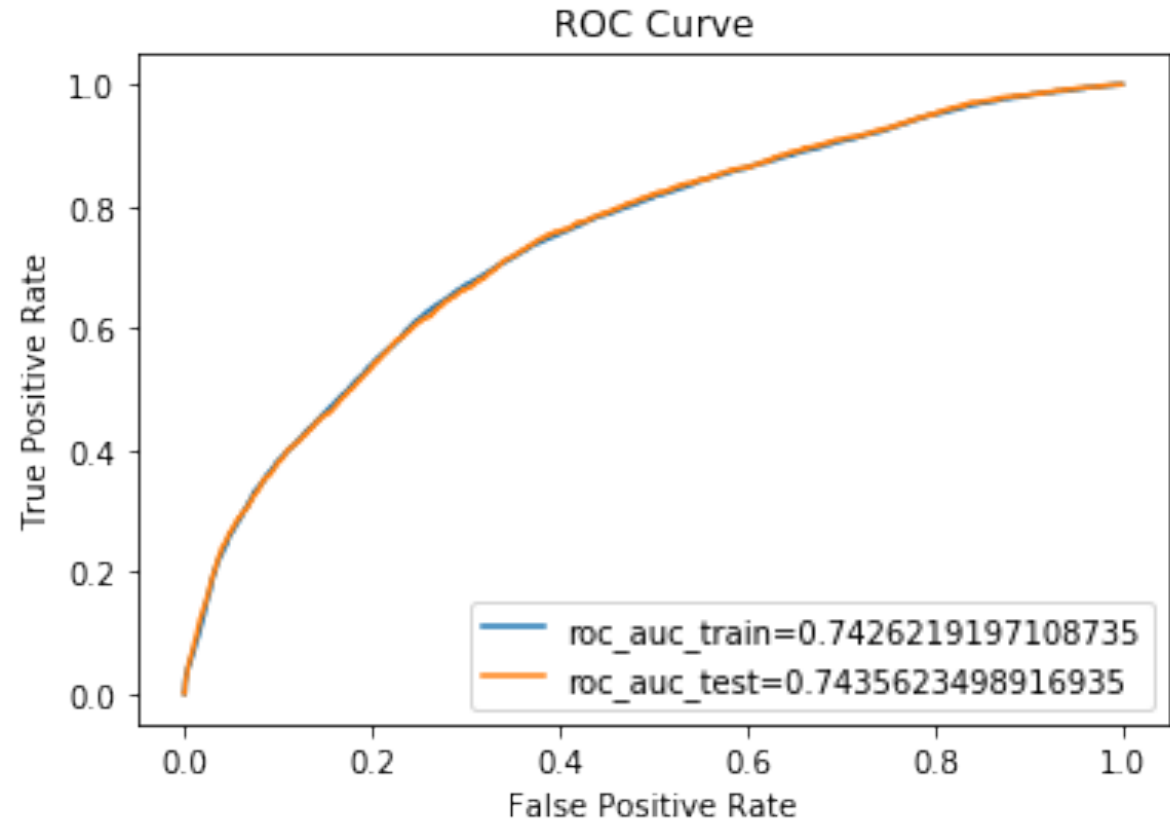
Modeling

- Model Selection and Tuning:
 - Logistic Regression
 - Decision Tree
 - Gradient Boosting Classifier
 - Random Forest Classifier

Modeling

Logistic Regression

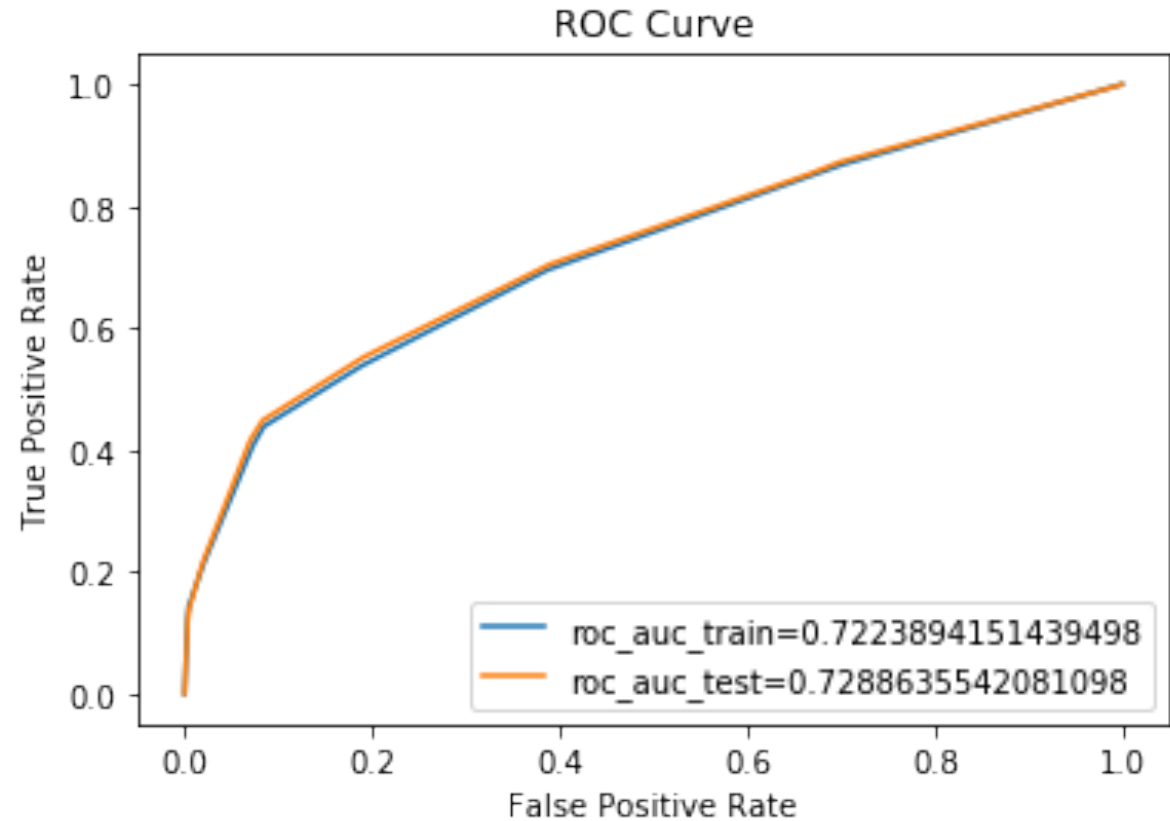
- AUC: 0.7597
- Precision: 0.08
- Recall: 0.67



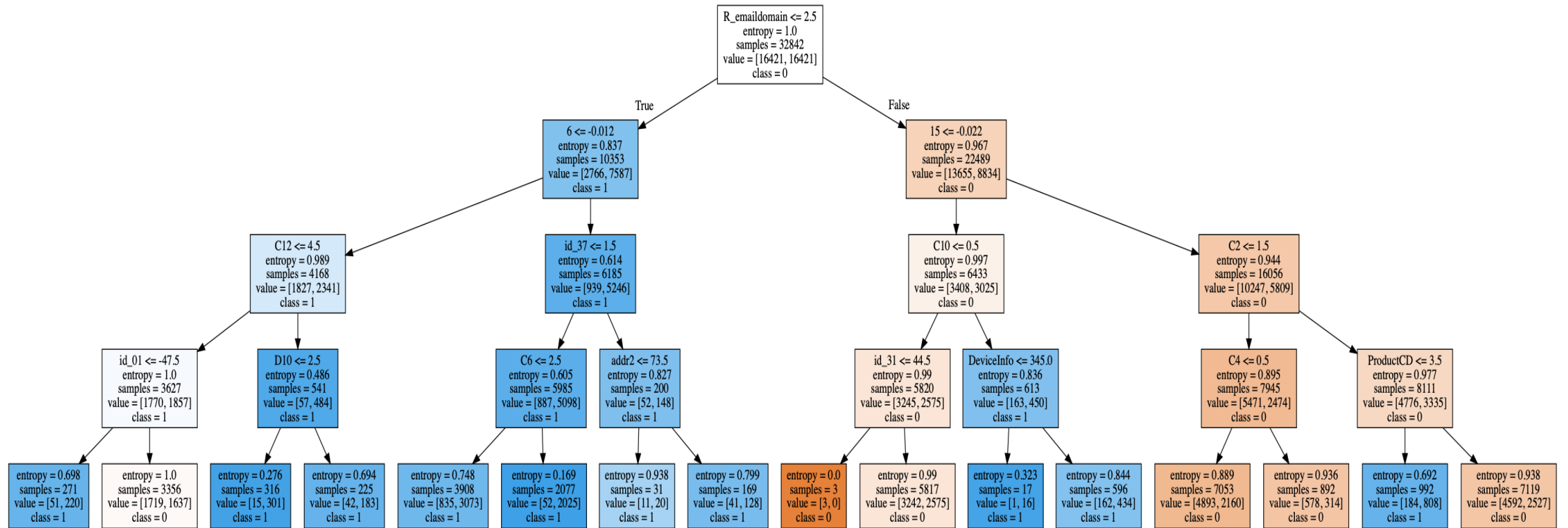
Modeling

Decision Tree

- AUC: 0.7288
- Precision: 0.17
- Recall: 0.45



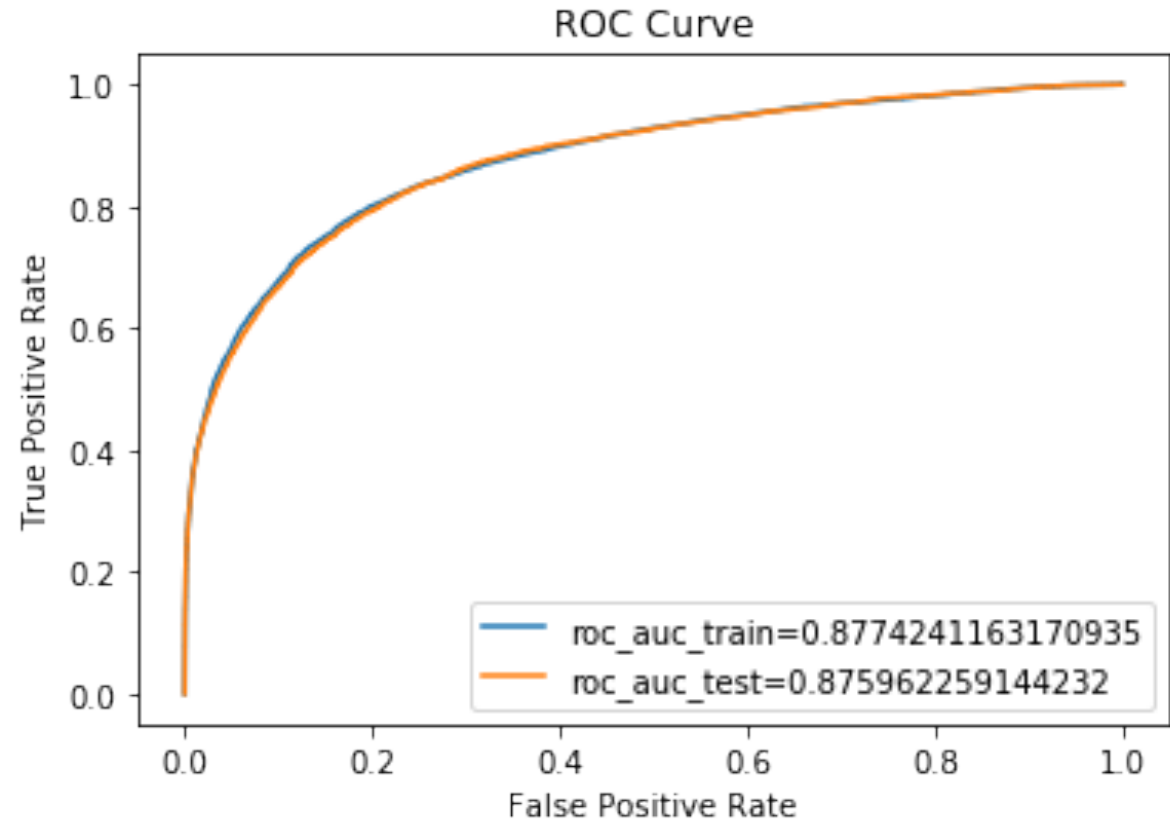
Modeling



Modeling

Gradient Boosting Classifier

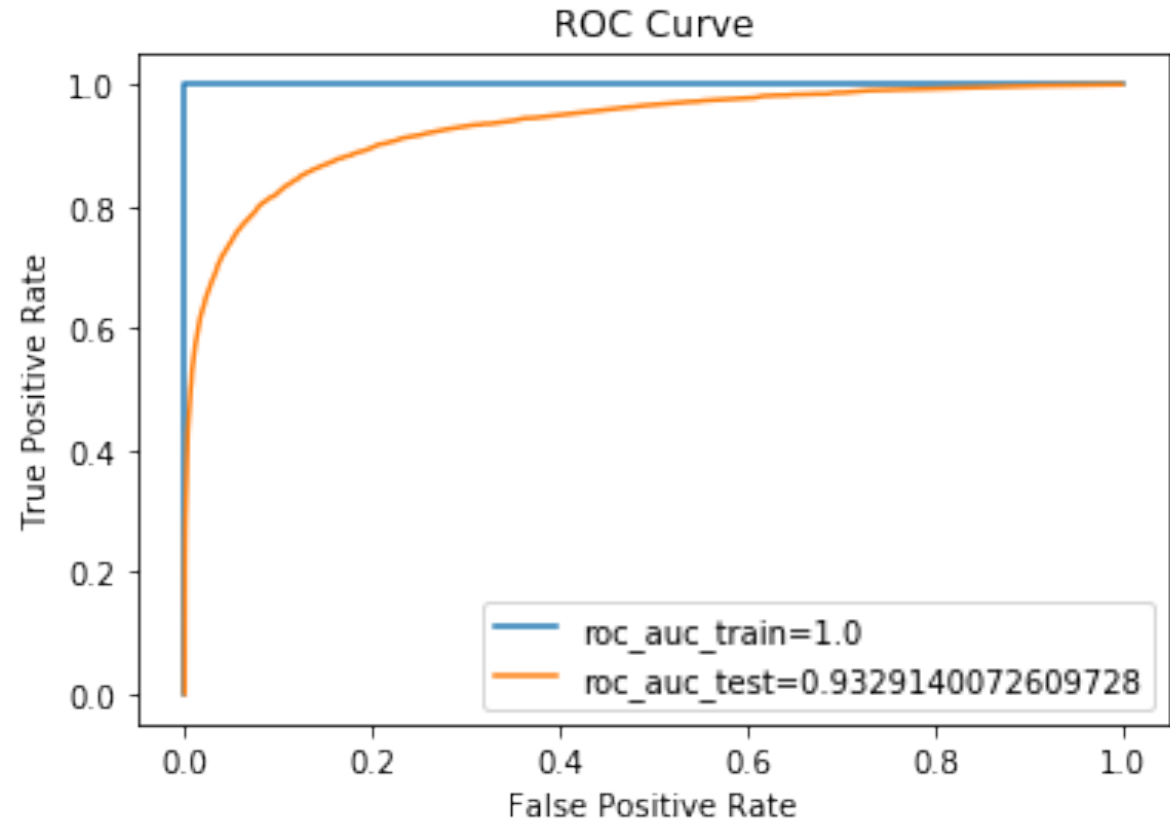
- AUC: 0.8759
- Precision: 0.14
- Recall: 0.77



Modeling

Random Forest (300 trees)

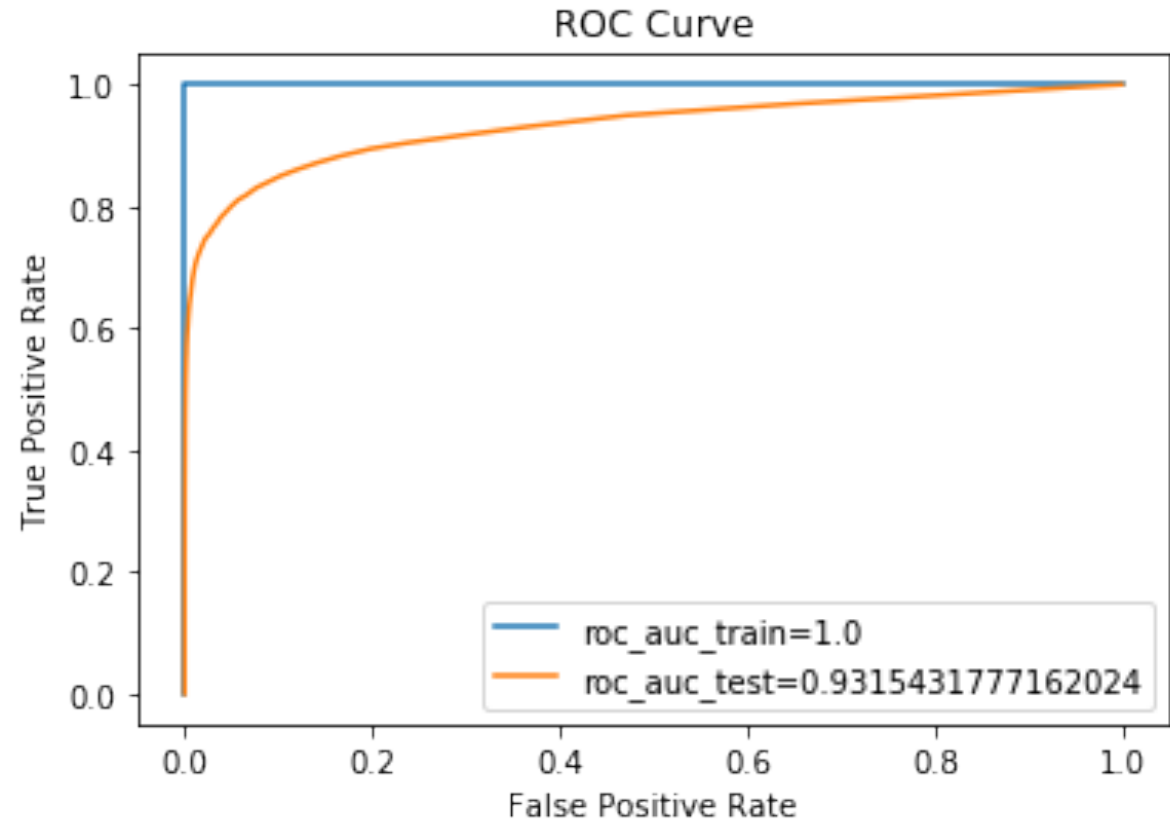
- AUC: 0.9329
- Precision: 0.21
- Recall: 0.84



Modeling

Random Forest (SMOTE)

- AUC: 0.9315
- Precision: 0.92
- Recall: 0.55



Modeling

Important Features (≥ 0.03):

- I (V-feature)
- Transaction Time Deltas
- Transaction Amount
- CardI
- CI3
- CI4
- I0 (V-feature)

Modeling

Model	AUC	Precision	Recall
Random Forest (300)	0.9327	0.21	0.84
Random Forest (SMOTE)	0.9315	0.92	0.55

Q&A