Challenge 2Model

Finding Fraud Faster

< Name: Xuhui Ying WFU ID: 06648543 >

Problem Statement

Since the pandemic, financial institutions have seen a 35% YoY increase in transaction fraud. Suppose I have just been invited to interview with the machine learning team at a large financial institution. Their job is to find fraud, waste, and abuse in the payment stream. I have been presented with a sample dataset of transactions and a holdout set, and my job is to walk through your process of exploring the data, calculating fraud rate, building three models, and evaluating them. My target variable is called EVENT_LABEL and contains a label "legit" or "fraud". In confusion matrixs, a false positive ratio is the probability of falsely rejecting the null hypothesis for a particular test. It can be calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events (regardless of classification). The reason they are trying to detect fraud at 5% FPR is a good way to prevent legit individuals and institutions be regarded as fraud.

Executive Summary

1. When the number of days since the account was created, the USD $ value of the transaction, the measure of the historic USD $ amount used to purchase goods and services increases, the fraud rate will also increase. When the adjustment of USD $ value to the transaction increases, the fraud rate decreases.

2. If Card Verification Value (CVV) is A, J, or P, the fraud rate must be 100%. If the code for the transaction environment is A, P, or Y, then the fraud rate must be 100%

3. Email domain and postal billing code are not important predictors to detect fraud because of the close to 0.5 area under the curve.

4. If the firm wants to operate at a 5% false positive rate, the score threshold should be 0.083. At this threshold, the recall should be 0.525; the precision should be 0.970.

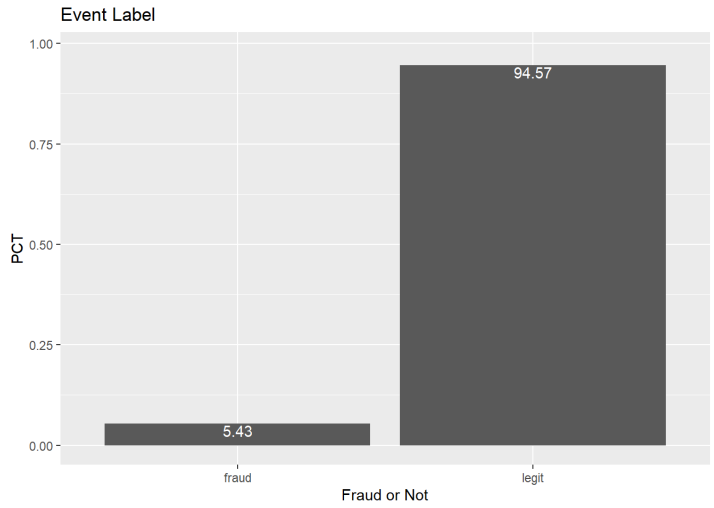
Recommendation

1. The financial institution should pay close attention to large value of USD transaction and large the measure of the historic USD $ amount used to purchase goods and services to prevent the fraud or reduce the fraud rate.

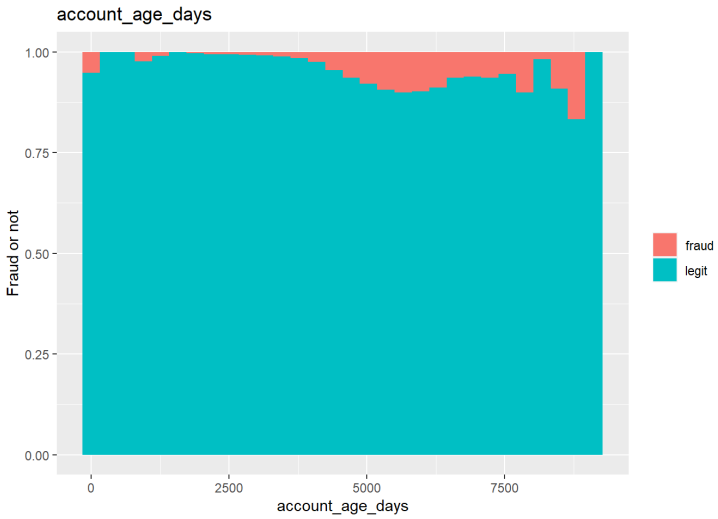
2. The financial institution should pay close attention to Card Verification Value (CVV) which is A, J, or P, and the code for the transaction environment which is A, P, or Y, because they are likely to fraud.

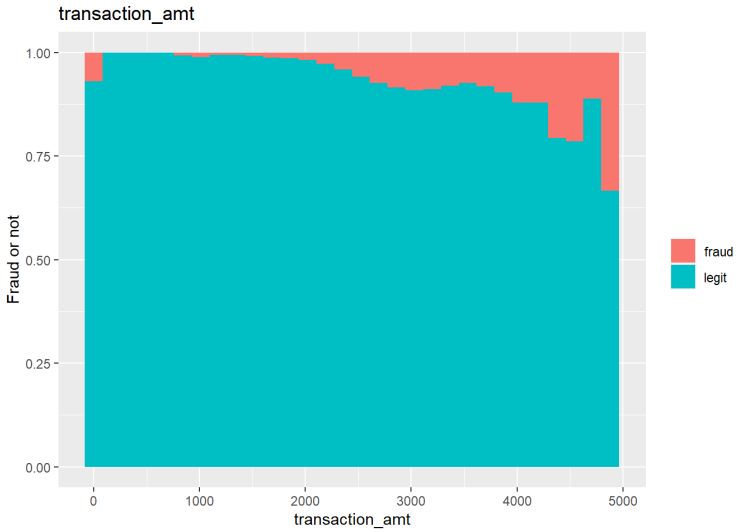
Analysis

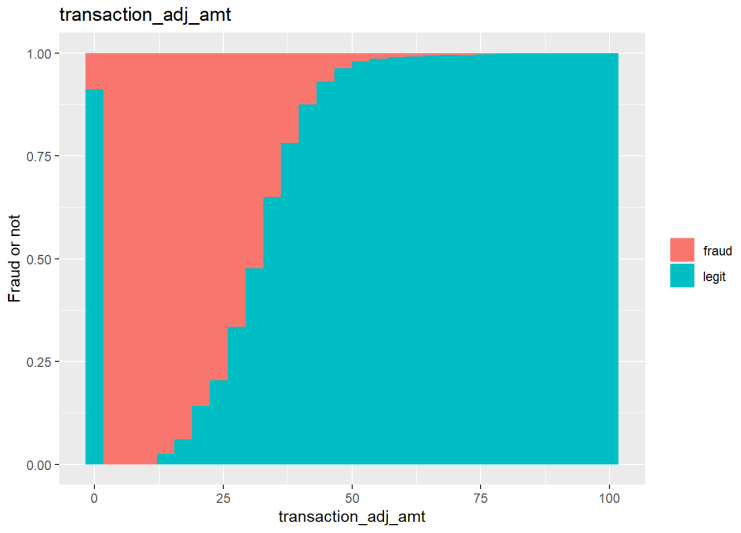
Explore Target

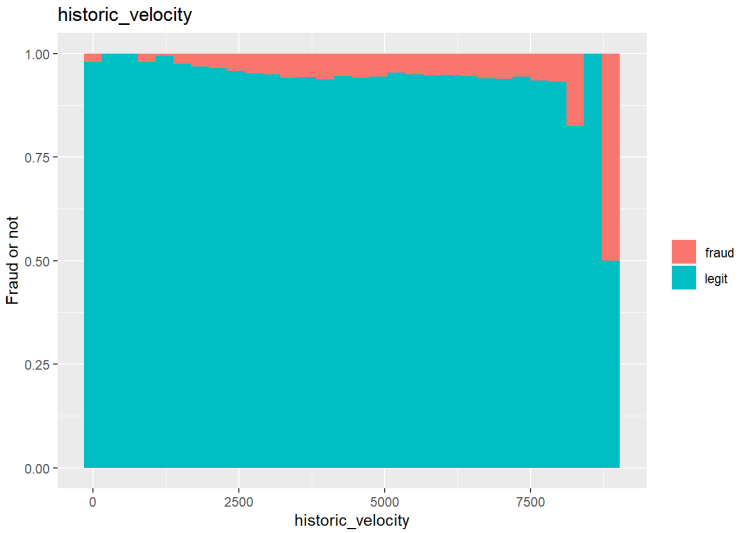


Explore numerics

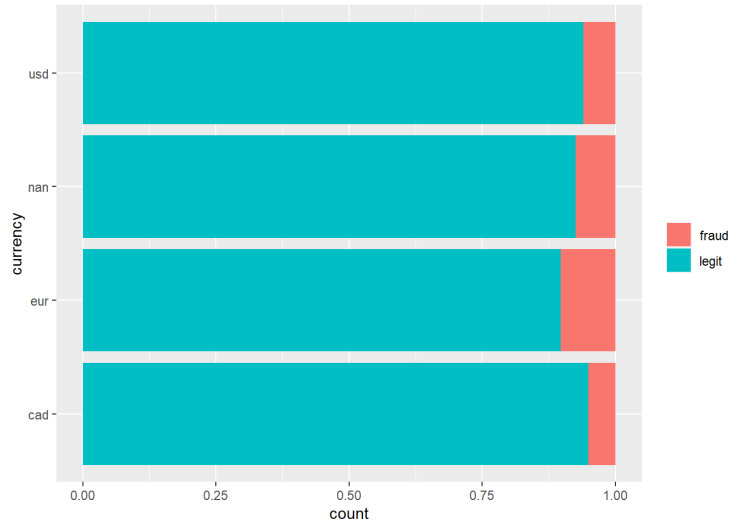


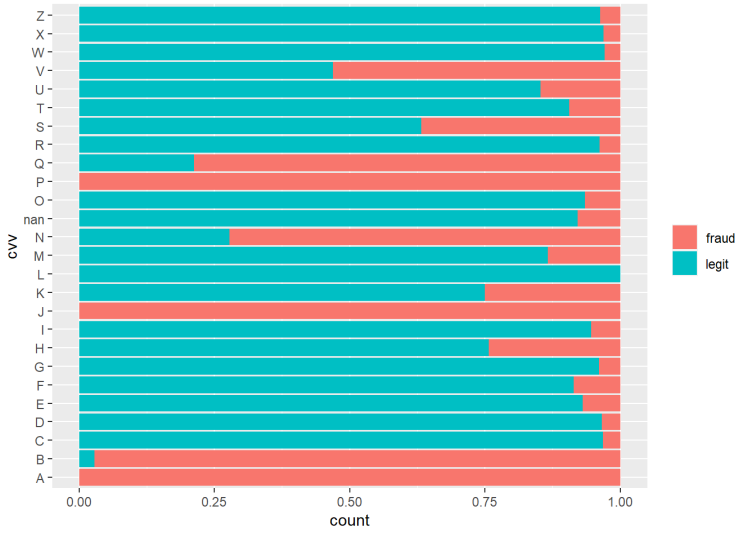

transaction_amt
Fraud or not
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
transaction_amt
fraud
legit

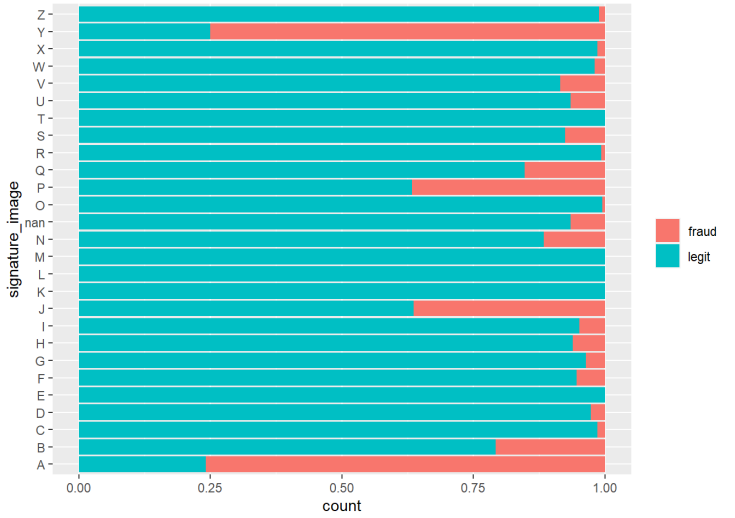

transaction_adj_amt
Fraud or not
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
transaction_adj_amt
fraud
legit

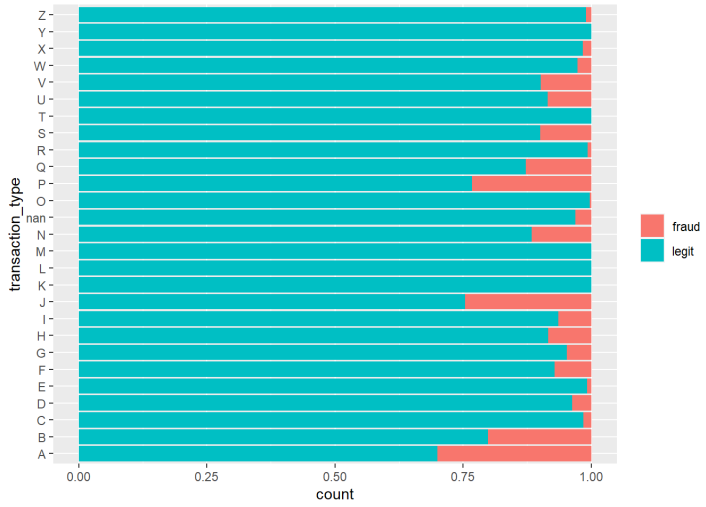

historic_velocity
Fraud or not
1.00
0.75
0.50
0.25
0.00
0
2500
5000
7500
historic_velocity
fraud
legit

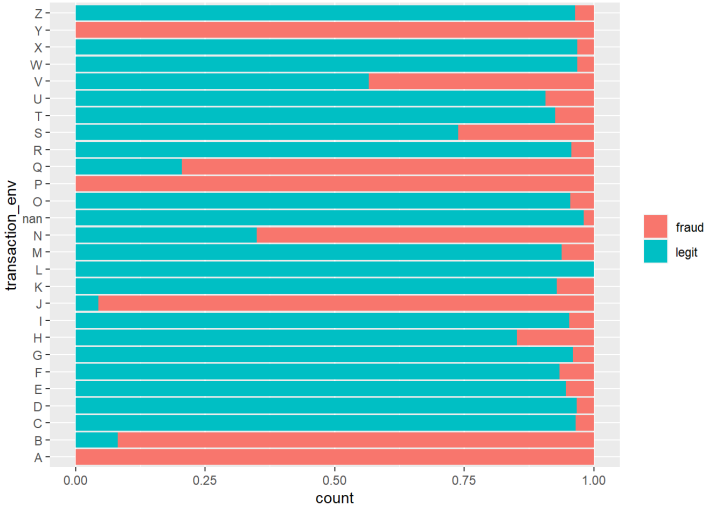
Explore character variables











Methodology

1. Data partitioning
    - Split the data into 70/30 train/test split using random sampling
2. Data preprocessing
    - Formula
        i. churn ~ account_age_days + transaction_amt + transaction_adj_amt + historic_velocity + currency + cvv + signature_image + transaction_type + transaction_env + billing_state
    - Numeric Predictor Pre-Processing
        i. Replaced missing numeric variables with median
        ii. Use an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time
    - Categorical Predictor Pre-Processing

    i. Replaced missing categorical variables with “unknown”
    ii. Dummy encoded categories with 1s and 0s
3. Model specification
    - Model 1: Decision Tree (Trees = 10, min_n = 10, importance="impurity")
    - Model 2: Decision Tree (Trees = 10, min_n = 10, importance=" permutation")
    - Model 3: Decision Tree (Trees = 1200, min_n = 10, importance="impurity")

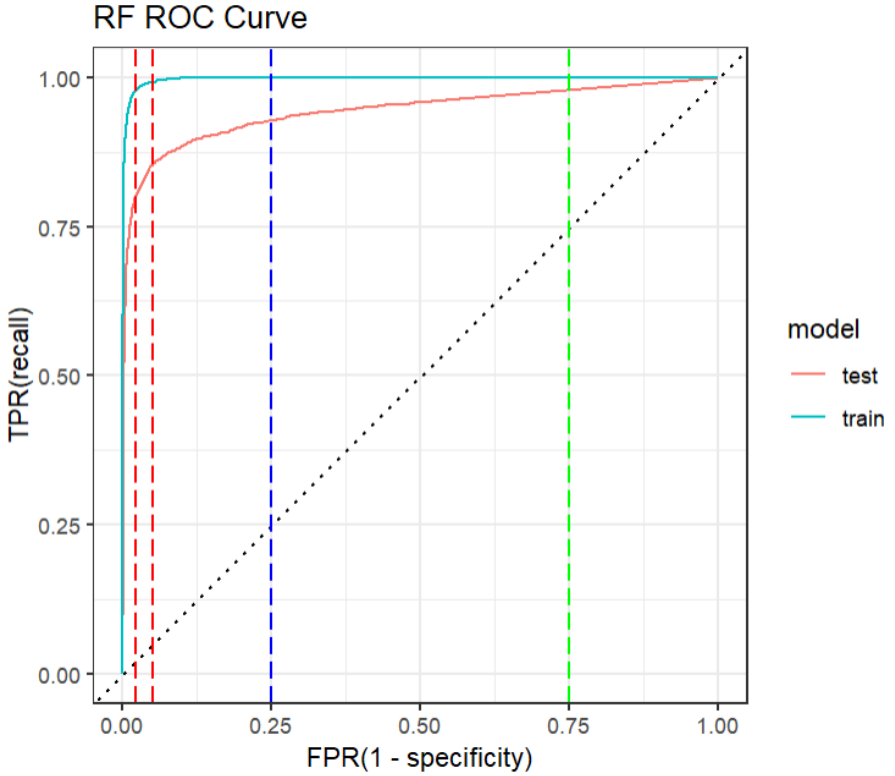
Model Analysis

<tables comparing your model performance >

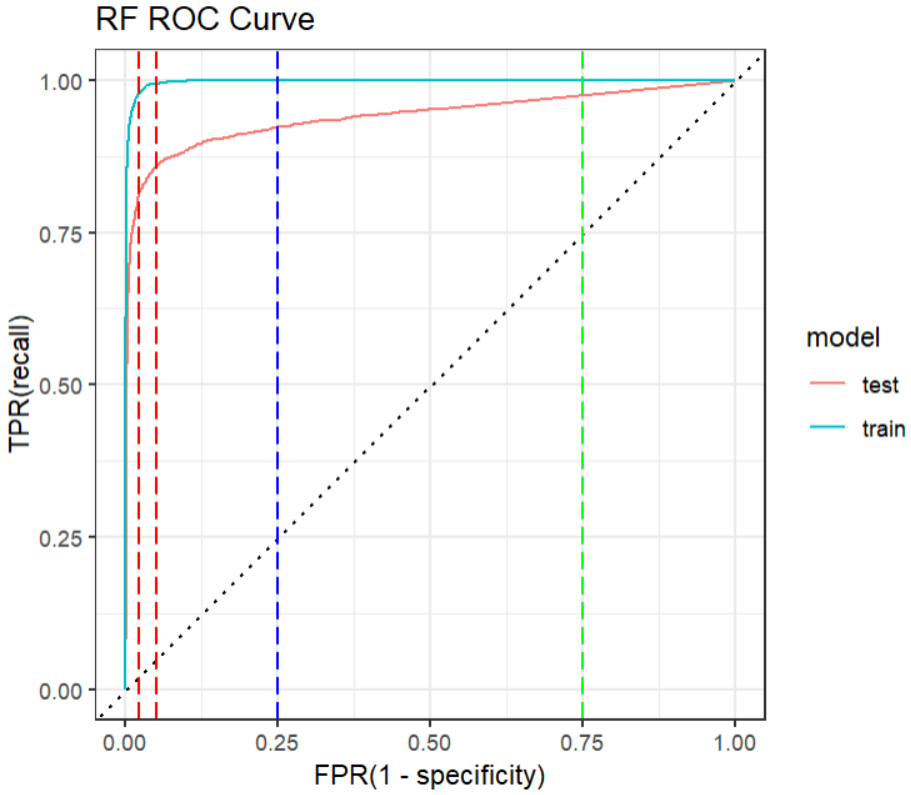
Model	Partition	Accuracy	AUC	Precision	Recall
Model 1	train	0.9867	0.9978	0.9850	0.7685
Model 2	train	0.9867	0.9980	0.9850	0.7685
Model 3	train	0.9878	0.9989	0.9913	0.7844
Model	Partition	Accuracy	AUC	Precision	Recall
Model 1	test	0.9746	0.9442	0.9031	0.5873
Model 2	test	0.9755	0.9414	0.9111	0.5998
Model 3	test	0.9770	0.9448	0.9246	0.6198

ROC Chart by Model

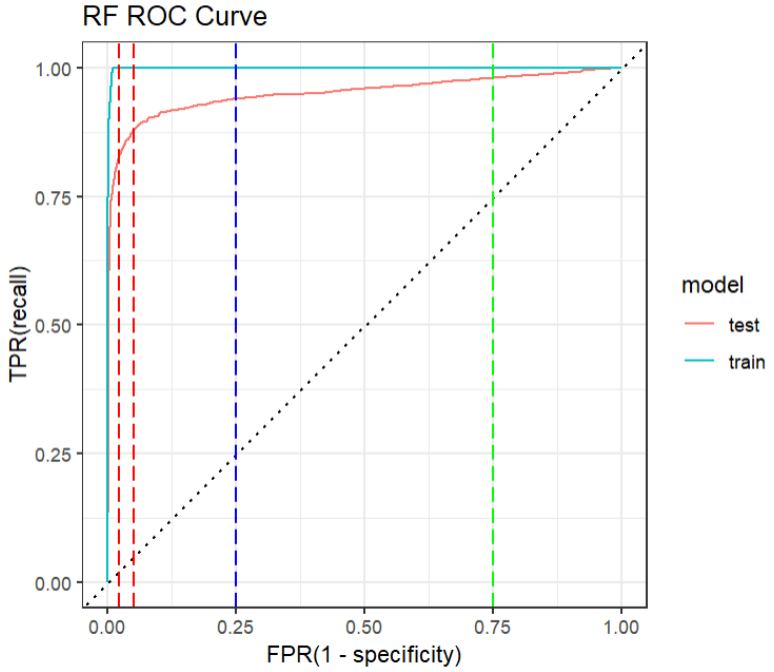
- Model 1: Decision Tree (Trees = 10, min_n = 10, importance="impurity")



- Model 2: Decision Tree (Trees = 10, min_n = 10, importance=" permutation")

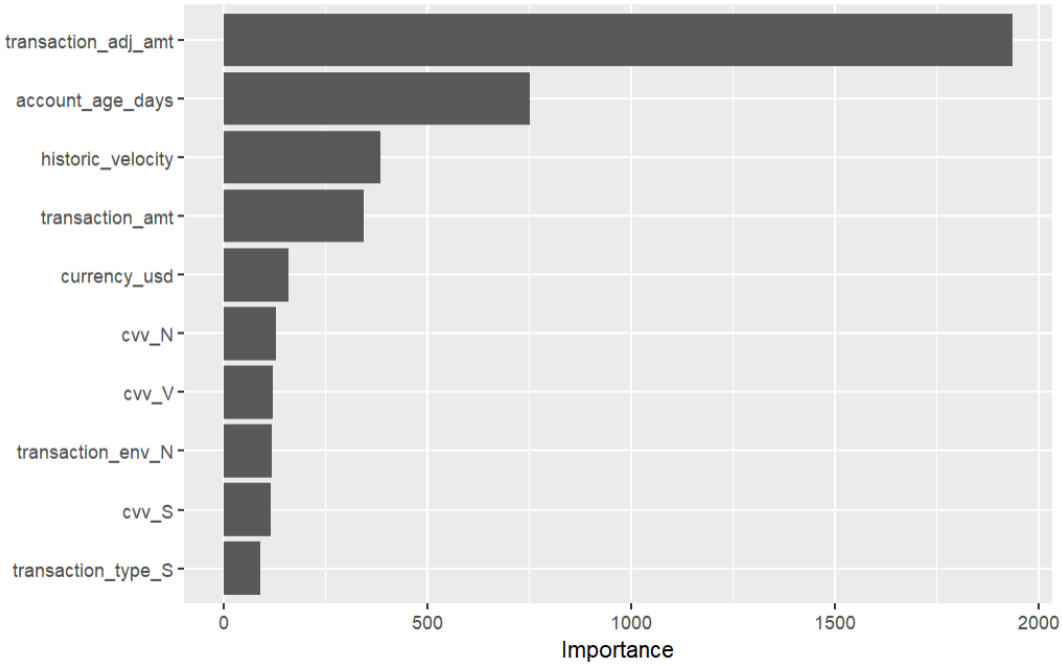


- Model 3: Decision Tree (Trees = 1200, min_n = 10, importance="impurity")

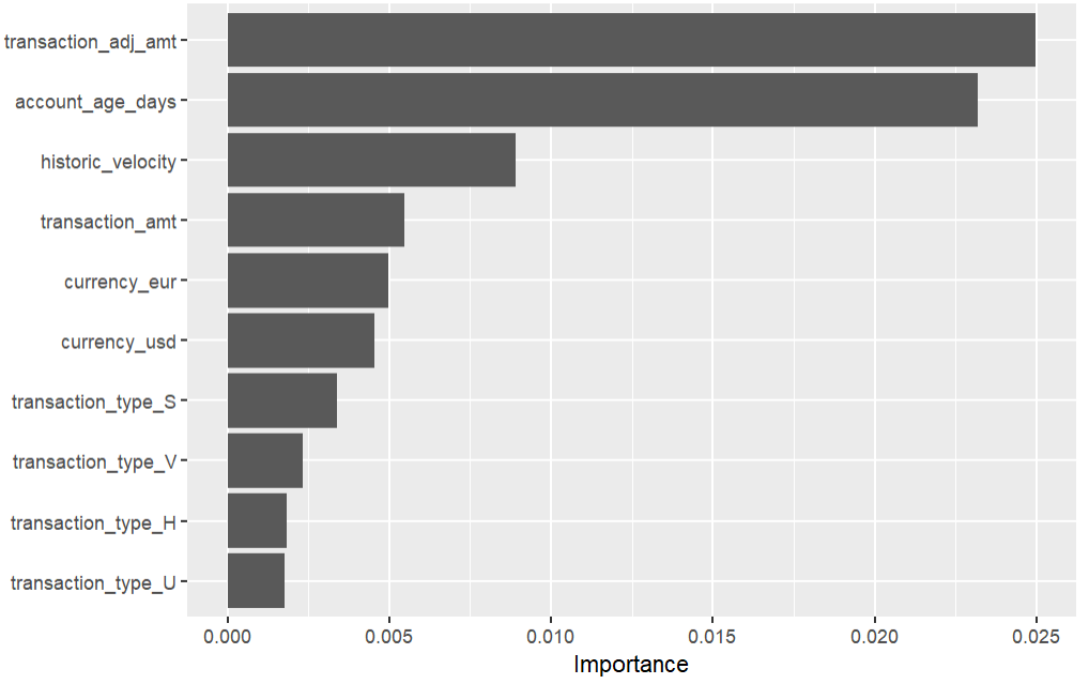


Feature Importance by Model

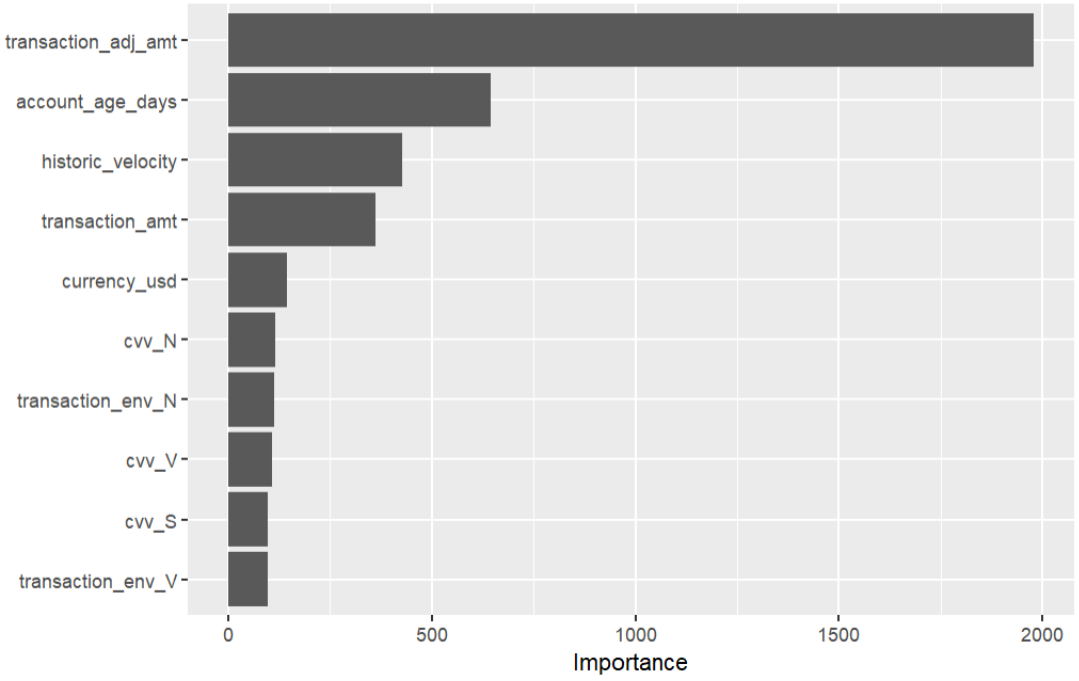
- Model 1: Decision Tree (Trees = 10, min_n = 10, importance="impurity")



- Model 2: Decision Tree (Trees = 10, min_n = 10, importance=" permutation")



- Model 3: Decision Tree (Trees = 1200, min_n = 10, importance="impurity")



Selected Model Operating Ranges

fpr	threshold	tpr	Precision	Recall
0.00	Inf	0.469	N/A	N/A
0.01	0.242	0.918	0.851	0.918
0.02	0.161	0.957	0.732	0.957
0.03	0.123	0.963	0.646	0.963
0.04	0.099	0.968	0.580	0.968
0.05	0.083	0.970	0.525	0.970
0.06	0.072	0.972	0.481	0.972
0.07	0.063	0.973	0.441	0.973
0.08	0.057	0.976	0.411	0.976
0.09	0.051	0.976	0.380	0.976
0.10	0.047	0.977	0.356	0.977

Operational Business Rules w. Expected Performance (Precision & Recall)

The above table shows that the threshold should be set at 0.083 to meet a 5% false positive rate to make the business function well. A threshold of 0.083 means that any transaction with a predicted fraud rate higher than 0.083 should be considered fraud. The precision at this threshold is 0.525, which means that about 52.5% of the predicted fraud is actual fraud. The recall at this threshold is 0.970, which means that 97.0% of the actual fraud was identified correctly.

Kaggle Submission

Kaggle Name: Eagle Xuhui Ying

Kaggle reported score: 0.95201

Kaggle reported position at time of submission: #10

(Note: this will change as others post)

https://www.kaggle.com/competitions/challenge-2-fraud-detection-2022/leaderboard