Finding Fraud Faster

< Emma Wang, Forest House>

Executive Summary

We built three models to identify fraud and decided the random forest model with 100 trees is better based on model evaluations. This model has an AUC of 0.946 (further explanation is shown below).

At the assigned 5% false positive rate, this model has a precision of 0.503 and a recall of 0.872, which means that half of those cases predicted as fraud are accurate and of all the fraudulent transactions, this model can successfully detect 87% of them.

Based on our findings, the financial institution should especially observe adjusted transaction amounts and consider closing certain types of CVVs and transaction environments.

Problem Statement

Business Background

We're working for a large financial institution whose job is to find fraud, waste, and abuse in the payment stream. Using provided historical transaction data, we created machine-learning models to identify fraud.

The 5% false positive rate

In this case, the false positive rate is assigned at 5%, under which they are trying to detect fraud. What does it mean operate at a 5% false positive rate or why did they choose this rate? They choose this rate because they can stand the consequence of mistaking 5% of the legit cases as fraud.

To do this, they will lower the threshold so more cases will be easier interpreted as fraud. By doing this, more fraud will be detected while more legit cases will be mistaken as

fraud, too. However, the price of assigning a legit transaction as fraud is lower than the price of interpreting a fraud as legit. Therefore, they are willing to mistake more legit cases as fraud in exchange for finding more actual fraudulent cases.

After comprehensive consideration of the prices of false positives and false negatives, the financial institution set the false positive rate at 5%, which is a result of leveraging the trade-off between precision and recall (explained as follows).

Introduction of Random Forest

Random Forest selects random predictors and a random sample of data and aggregates those samples' classification results to make a final classification decision.

Evaluating Metrics

- Area Under the ROC Curve
    - The Area Under the Curve (AUC), which represents the area under the ROC curve, is often used when comparing models. For an effective model, the AUC should be higher than 0.5, which means the model is better than random guessing. The higher AUC, the better the model.
- Precision
    - Precision is TP/(TP+FP), Where TP = True Positives, FP = False Positive. In other words, precision measures the accuracy of predicted fraud, i.e., whether those predicted frauds are actual frauds.
- Recall
    - Recall is TP/(TP+FN), Where TP = True Positives, FN = False Negatives. In other words, recall measures the percentage of frauds that are successfully identified.

Methodology

1. Data partitioning
    - Split the data into 70/30 train/test split using random sampling
2. Data preprocessing
    - Formula

    - event_label ~ account_age_days + transaction_amt + transaction_adj_amt + historic_velocity + billing_state + currency + cvv + signature_image + transaction_type + transaction_env
- Numeric Predictor Pre-Processing
    - Replaced missing numeric variables with the median
    - Centered and scaled numeric predictors to have a mean of 0 and standard deviation of 1
- Categorical Predictor Pre-Processing
    - Replaced missing categorical variables with unknown
    - Pool all levels in each categorical variable with a rate of occurrence less than 1% into an "other" level.
    - Dummy encoded categories with 1s and 0s

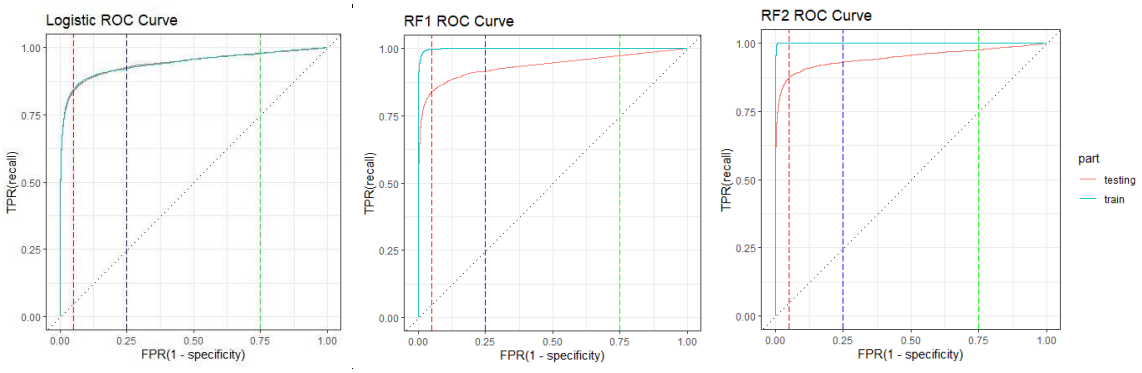
3. Model specification
    - Train a Logistic Regression model
    - Train a Random Forest model with tree=10
    - Train a Random Forest model with tree=100

Model Comparison

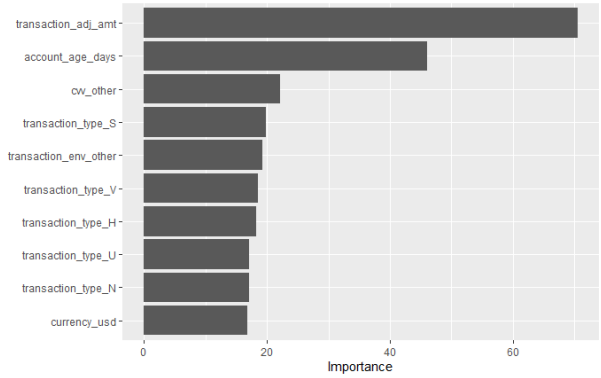
Comparison Table

Model	Partition	AUC	Precision	Recall
Logistic Regression	train	0.941	0.861	0.606
Random Forest (10)	train	0.999	0.987	0.795
Random Forest (100)	train	1.00	0.994	0.809
Logistic Regression	test	0.941	0.863	0.609
Random Forest (10)	test	0.936	0.905	0.591
Random Forest (100)	test	0.946	0.930	0.611

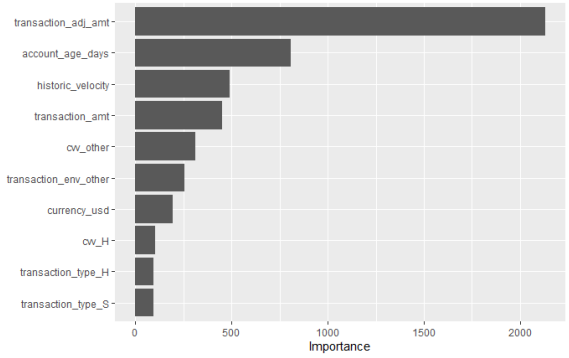
ROC Chart by Model



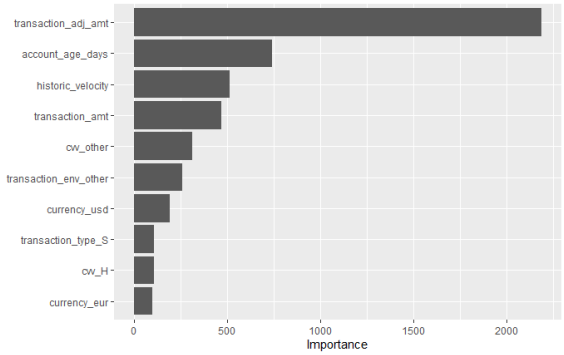
Feature Importance by Model



The top 10 most important variables generated by the logistic regression model are shown above.



The top 10 most important variables generated by the random forest model with 10 trees are shown above.



The top 10 most important variables generated by the random forest model with 100 trees are shown above.

Combining the three results, transaction-adjusted amount, account age days, historic velocity, and some CVV levels and transaction environment levels are most useful.

Model Selection

Based on the information above, we believe that model3, i.e., the random forest model with trees=1000 is the best model.

First, the test part's AUC is the highest among the three models, which means it can classify fraud and legit cases better.

Second, the test part's precision, and recall are also the highest among the three models, meaning this model can capture more actual fraud successfully and the predicted fraud of this model has higher accuracy.

Third, the AUC of this model's test and training dataset is about the same, which means that there are no overfitting problems. Therefore, model3 is the best model.

Analysis

Target Analysis

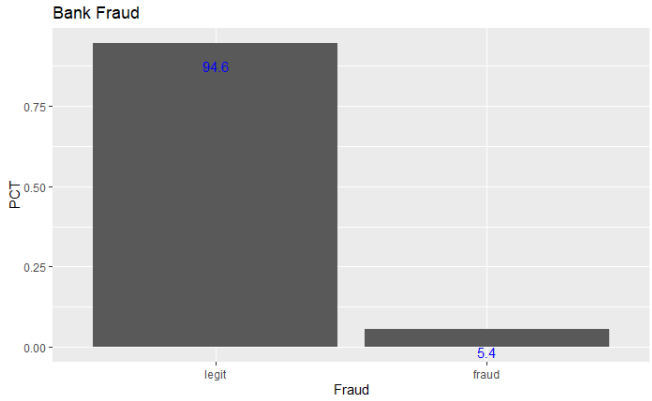


Chart-1

As the chart shows above, nearly 5.4% of the transactions are fraudulent.

Numeric Variables Analysis

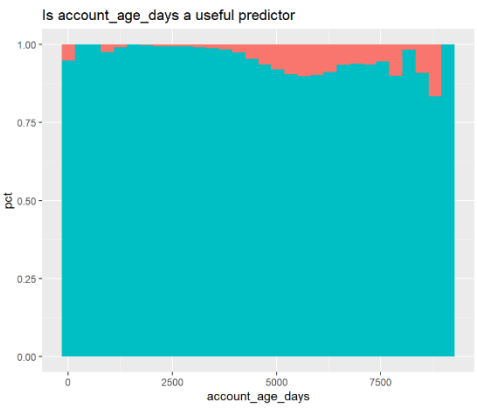


Chart-2

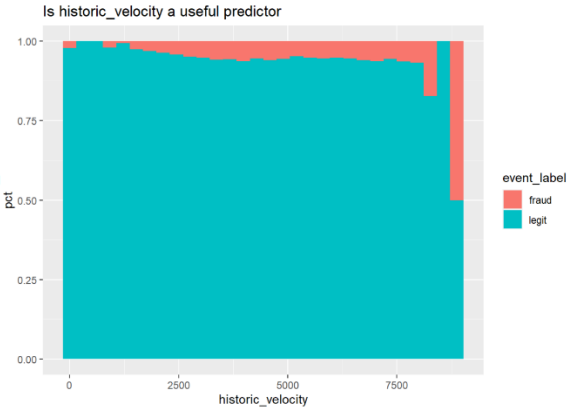


Chart-3

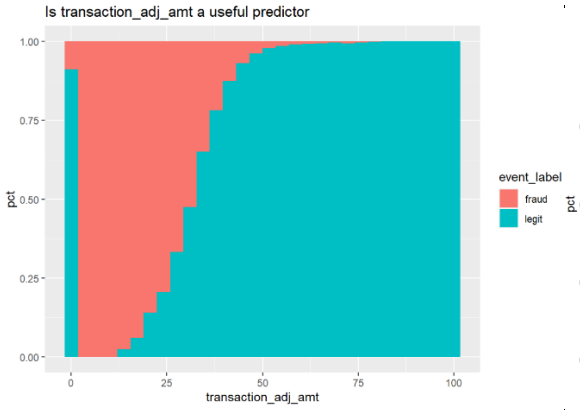


Chart-4

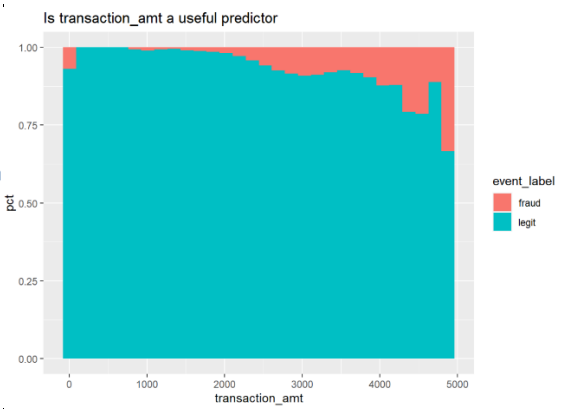


Chart-5

We picked four numeric variables from the top 10 important variables in the Random Forest model with 100 trees. As the charts show above, account age, transaction amount, transaction adjusted amount, and historic velocity can be useful when predicting fraud, as the distribution of the percentage of fraud is uneven among different amounts.

Categorical Variable Analysis

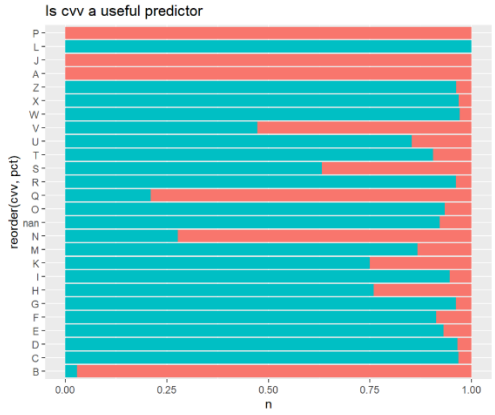


Chart-6

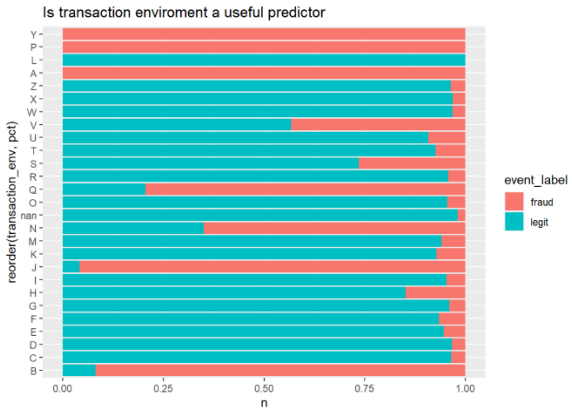


Chart-7

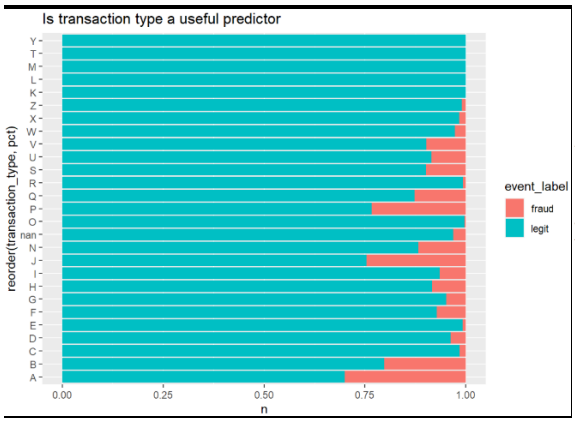


Chart-8

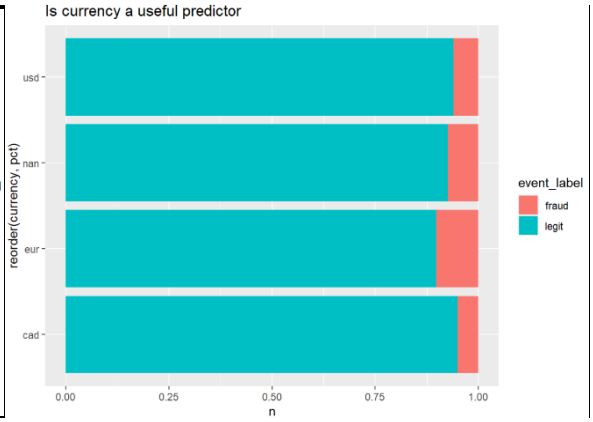


Chart-9

We picked four categorical variables from the top 10 important variables in the Random Forest model with 100 trees. As the charts show above, cvv, transaction type, transaction environment, and currency can be useful when predicting fraud, as the distribution of the percentage of fraud is uneven among different levels.

Insights

- **The most important variable is the transaction-adjusted amount.** As the distribution of the percentage of fraud amount under different transaction-adjusted amounts is the most uneven one compared to other variables. This is also proved by the three models' important features result.
- **Neither email domain nor billing postal code is an important predictor.** To analyze this, we conducted extra research by putting only these two variables in a regression model, separately, and studying different levels' p-value. According to the results, insignificant levels in the email domain account for 94% (111/118) and 95% (36/38) in billing postal.
- **A higher historic velocity or transaction amount indicates a higher probability of fraud.** As shown in chart-3 and chart-5, the percentage of fraud increased significantly when historical velocity exceeds approximately $8700 and transaction amount exceeds approximately $4800.
- **Euros have a higher chance to be used in fraudulent activities.** As shown in chart 9, euros have the highest fraud percentage among all three kinds of currency. Nearly 20% of the transactions with euros are frauds.

Selected Model Operating Ranges

fpr	threshold	tpr	Precision	Recall
0	1	0.368	-	-
0.01	0.330	0.746	0.826	0.747
0.02	0.217	0.813	0.707	0.813
0.03	0.164	0.842	0.619	0.842
0.04	0.132	0.861	0.559	0.860
0.05	0.110	0.872	0.503	0.872
0.06	0.094	0.881	0.461	0.881
0.07	0.083	0.886	0.422	0.886
0.08	0.073	0.891	0.391	0.892
0.09	0.065	0.894	0.364	0.894
0.10	0.059	0.901	0.343	0.900

Based on our best-performing model, the following rule can satisfy a 5% false positive rate: *If the probability of fraud is bigger than or equal to 0.117, then it is defined as fraud, otherwise, it will be defined as legit.* The recall and precision at this threshold are 0.503 and 0.872, respectively. This means that half of those cases predicted as fraud are accurate and of all the fraudulent transactions, this model can successfully detect 87% of them. A 5% FPR means that 5% of the legit transactions will be mistakenly predicted as fraud.

Recommendations

1. Use transaction-adjusted amount as a primary metric to detect fraud, as it is the most important variable according to the result of all three models. Especially keep an eye out for transactions with an adjusted amount lower than 25$, according to chart-4 the probability of fraud is nearly 100% when the adjusted amount is lower than 12.5$.
2. Investigate or close off CVVs coded as P, J, A, B, and transaction environment coded as Y, P, A, J, B. According to chart-6 and chart-7, more than 90% of transactions made under such CVVs or environments are fraudulent. There might be bugs existing in those systems or procedures. If the investigation turns back no methods to fix it, the financial situation should consider shutting them down.

Kaggle Submission:

Kaggle Name: Emma Wang (Jiawen)

Kaggle reported a score: of 0.95125

Kaggle reported the position at the time of submission: #12