Finding Fraud Faster

< Yi Ding >

# Executive Summary

Key Findings

* Email domain and billing postal code both have a p-value greater than 0.05, which is considered statistically insignificant in determining fraud status, meaning that they cannot effectively improve the overall performance of the model.
* According to the variable importance plot, the adjustment USD $ value to the transaction, number of days since the account was created, the USD $ value of the transaction, original currency code, transaction type and card verification value are the most useful indicators in predicting fraud.
* If the 2nd random forest model operates at a 5% false positive rate, where the threshold is 0.362, the precision would be 0.4977 and the recall would yield to 0.8521.

## Analysis

Metrics

We will evaluate the models using FPR (false positive rate), the probability of falsely identifying bank fraud. The false positive rate is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events. Formally, FPR has the following definition:

Where FP = False Positives, TN = True Negatives

Chart

Description automatically generated

The chart above shows 94.57% of legit transactions, while 6785 of cases were fraud. In this project, we will focus on the fraud label as our target.

Table

Description automatically generatedTable

Description automatically generatedTable

Description automatically generated

The chart above is from logistic regression, showing the usefulness of each variable from a statistical perspective. In general, a useful variable should be statistically significant (p-value < 0.05). Thus, we filtered out variables which are not significant, such as billing\_postal, card\_bin, days\_since\_last\_logon, initial\_amount, email\_domain, and locale, as they are not good predictors.

Moreover, we also dropped variables with many unique values, such as event\_id, ip\_address, applicant\_name, merchant\_id and billing\_address, since we want to predict a more general pattern of the target (fraud), unique identifiers won’t be helpful in that case.

## Methodology

1. Data partitioning

* Split the data into 70/30 train/test split using random sampling

1. Data preprocessing

* Formula
  1. event\_label ~ billing\_state + currency + cvv + transaction\_type + transaction\_env + account\_age\_days + transaction\_amt + transaction\_adj\_amt
* Numeric Predictor Pre-Processing
  1. Replaced missing numeric variables with median
* Categorical Predictor Pre-Processing
  1. Replaced missing categorical variables with “unknown”
  2. Removed rows to make the occurrence of levels in a specific factor level equal
  3. Pooled infrequently occurring values into an "other" category
  4. Dummy encoded categories with 1s and 0s

1. Model specification

* Random Forest Model 1 (trees = 100, min\_n = 20)
* Random Forest Model 2 (trees = 200, min\_n = 10)
* Logistic Regression Model

## Recommendations

* The firm should not use email domain and billing postal code in predicting fraud as they are not effective indicators to determine if a transaction is fraud or not.
* The firm should pay more attention to transaction type, especially for C, R, X, Z, since transactions under these types are more likely to be fraud.
* A random forest model creates a collection of decision trees and use these trees to benchmark each other and chooses the most common prediction as the final result. The benefit of using random forest is that even if one of its trees makes an error, more other trees will likely avoid making the same error, resulting in a more accurate prediction.
* The loss caused by a true fraud is significantly higher than that caused by a false fraud prediction. Thus, the firm needs to make certain compromises in false fraud predictions, to optimize its ability to identify true frauds. To achieve this, the firm is recommended to operate at an FPR below 6%, as the marginal return in accurate true fraud identification diminishes significantly beyond FPR at 6%.

## Model Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Partition | AUC | Precision | Recall |
| Rand Forest 1 | train | 0.9883936 | 0.6930609 | 0.8680482 |
| Rand Forest 2 | train | 0.9927411 | 0.7050034 | 0.8879256 |
| Logistic Model | train | 0.9414706 | 0.5767610 | 0.8103193 |
|  |  |  |  |  |
| Model | Partition | AUC | Precision | Recall |
| Rand Forest 1 | test | 0.9411479 | 0.6461357 | 0.8010700 |
| Rand Forest 2 | test | 0.9416318 | 0.6503968 | 0.7971790 |
| Logistic Model | test | 0.9405226 | 0.5750521 | 0.8049611 |

Based on the chart above, random forest model 2 performs the best as it has the highest AUC and precision score with a slightly lower recall rate. An AUC = 0.9416 means that the model is 94.16% accurate in distinguishing the fraud and legit cases. A precision = 0.6504 tells that for all fraud cases we predicted, 65.04% of them are correctly identified. A recall = 0.7972 indicates that we can correctly identify 79.72% of all fraud cases that actually happened.

But except for the logistic regression model, the rest of two models prone to have overfitting issues due to the large gap between their training and testing dataset.

ROC Chart by Random Forest Model 2

An ROC curve entails how capable the model is at distinguishing between results. As ROC approaches 1, the model is more likely to predict a positive as positive, and a negative as negative. As we can see there is a gap between training and testing curves, which implies the model is overfitting, results in worse performance in testing part.

Chart, line chart

Description automatically generated

Feature Importance by Model

The graph below shows the top 10 important variables in random forest model 2, which are transaction\_adj\_amt, account\_age\_days, transaction\_amt, currency\_usd, currency\_cad, transaction\_env\_other, transaction\_type\_H, cvv\_other, currency\_eur, and transaction\_type\_S.

Chart

Description automatically generated

Selected Model Operating Ranges

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FPR** | **threshold** | **TPR** | **Precision** | **Recall** |
| 0 | Inf | 0.3258913 | - | - |
| 0.01 | 0.680 | 0.6872358 | 0.8085227 | 0.6921206 |
| 0.02 | 0.547 | 0.7764461 | 0.69569 | 0.7772374 |
| 0.03 | 0.463 | 0.8128883 | 0.6146755 | 0.8107977 |
| 0.04 | 0.406 | 0.8385769 | 0.5506876 | 0.8375486 |
| 0.05 | 0.362 | 0.8540663 | 0.4977273 | 0.8521401 |
| 0.06 | 0.324 | 0.8712231 | 0.4584293 | 0.8715953 |
| 0.07 | 0.294 | 0.8809489 | 0.4223154 | 0.8818093 |
| 0.08 | 0.269 | 0.8881335 | 0.3914629 | 0.8876459 |
| 0.09 | 0.247 | 0.8937030 | 0.365354 | 0.8934825 |

According to the firm’s requirements, the model should operate at a 5% false positive rate. The highlighted row in the chart reveals model’s expected performance metrics at the required rate. A 5% FPR results in an 85.4% TPR, and the threshold = 0.362. Following the requirement, the operating rule would be if .pred\_fraud ≥ 0.362, then it’s fraud, otherwise, it’s legit.

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

Precision is a percentage of true positive among all predicted positive cases, while recall is a percentage of true positive among all actual positive cases.

According to the operating range above, FPR increases as the threshold decreases, while precision decreases and recall increases. FPR is the percentage of predicting an actual negative case as positive in terms of total actual negative cases. In this project, FPR means, in total actual legit transactions, the percentage of incorrectly identified a legit transaction as a fraud. As FPR increases, the possibility of accurately predicting fraud (TPR) increases, but the precision would be lower since we allow more inaccurate predictions.

This is actually a tradeoff between precision and recall, if we want higher precision, we should lower the FPR (raise the threshold), the model will allow for more false negatives to slip by; if we want higher recall, then we should increase the FPR (lower the threshold), the model will allow for more false positives to slip by.

More specifically, as FPR increases, recall increases more less (recall increases the most when FPR increases from 0.01 to 0.02). Therefore, when setting the threshold, a balanced selection should be made to maximize the return of both precision and recall.