Challenge 2 Report
Finding Fraud Faster

Charlene Wei

1. Executive Summary

This project wants us to use some predictive models to predict which types of transfers are more likely to be scammed. Therefore, companies can help customers avoid losses by identifying these transfers.

During the exploratory analysis phase, 5% of the transfers in the dataset were fraudulent. At the same time, the graphs in the exploratory analysis indicate that several factors are related to encountering fraud.

In the Methods section, the three used models (decision trees and two random forests) are included. This section describes how to prepare and split the data into train and test parts, as well as define recipes and fit models

In the evaluation part, use accuracy, roc auc, precision and recall to evaluate the accuracy of the model.

In the evaluation and recommendation section, two variables that are concerned by the company, emaildomain and billingpostal, are discussed, and they are largely irrelevant to the target value. Also based on the analysis of Feature Importance, some useful predictors were found, such as: transaction_adj_amt, and account_age_days.

Based on the model, several recommendations are made.

- When trying to predict scams, we should not give too much consideration to the two variables email domain and billing postal.
- Banks should remind users who have opened accounts for a long time to pay attention to the security of transfers, either by mail or by phone.
- For transfers exceeding a certain limit, the bank should use multiple verification methods to ensure that the transfer is made by the person and to ensure the security of the transfer.
- We recommend that companies implement security measures for transfer environments where fraud is high.

2. Introduction

2.1. Challenge Overview

Transaction fraud rates at financial institutions have increased by 35% year-over-year since the pandemic. Their job is to spot fraud, waste, and abuse in the payment flow.

2.2. Problem Statement

This project aims to explore what factors are associated with ‘fraud’ by means of three machine learning models.

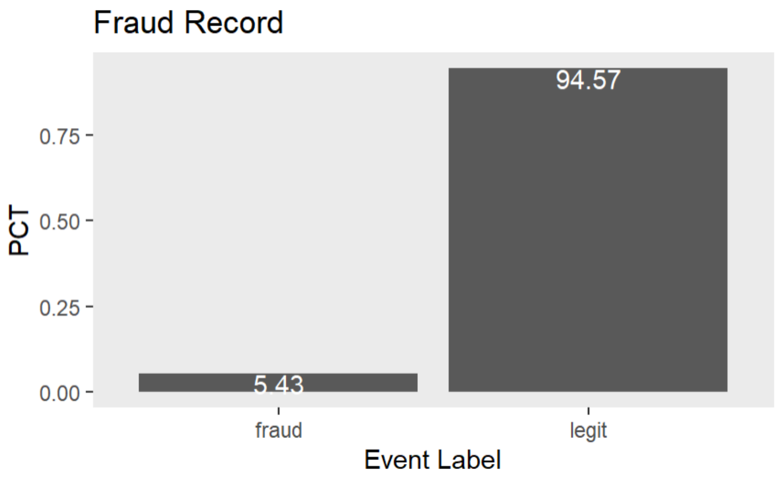
2.3. Data dictionary

Variable	Description
EVENT_ID	Transaction Identifier
account_age_days	number of days since the account was created
transaction_amt	the USD $ value of the transaction
transaction_adj_amt	the adjustment of USD $ value to the transaction
historic_velocity	The measure of the historic USD $ amount used to purchase goods and services
ip_address	ip address of the transactor
user_agent	user agent of the transactor
email_domain	email domain of the transactor
phone_number	phone number of the transactor
billing_city	billing city name
billing_postal	billing postal code
billing_state	billing state code
card_bin	first 6 digits of the credit card (determines the card type, issuing bank, debit/credit/prepaid)
currency	original currency code
cvv	Card Verification Value - the 3 digit number on back fo your card
signature_image	code for the signature
transaction_type	code for the transaction type
transaction_env	code for the transaction environment
EVENT_TIMESTAMP	timestamp when the transaction occurred
applicant_name	name of the transactor - ignore
billing_address	billing address of the card holder
merchant_id	merchant identifier
locale	browser locale
tranaction_initiate	code for type of transaction initiation
days_since_last_logon	days since last transaction initiated
inital_amount	amount of first transaction USD $

EVENT_LABEL	TARGET fraud / legit		

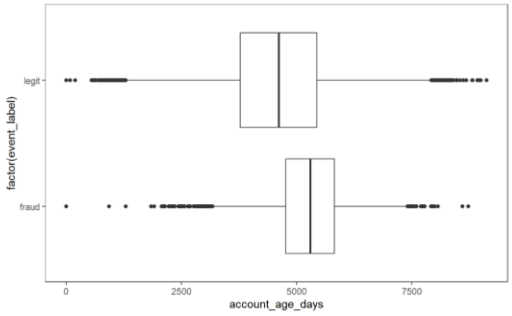
3. Exploratory analysis

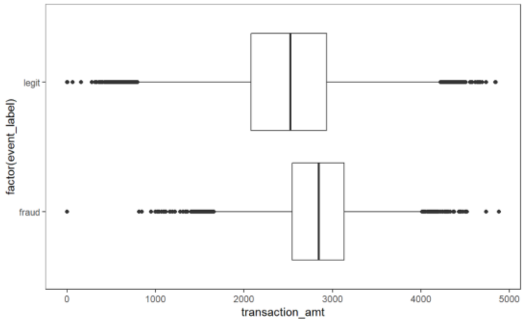
3.1. Exploring target variable



The event label target variable analysis reveals that over 5% of consumers had experienced fraud in the past.

3.2. Exploring numeric variables

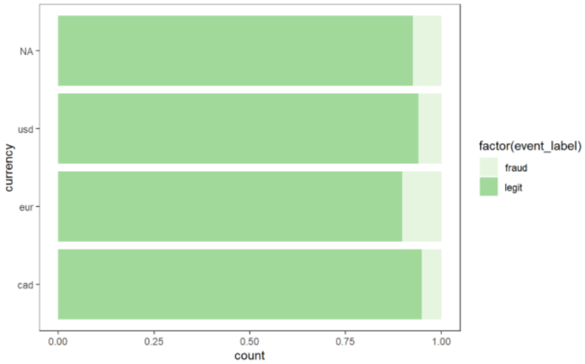


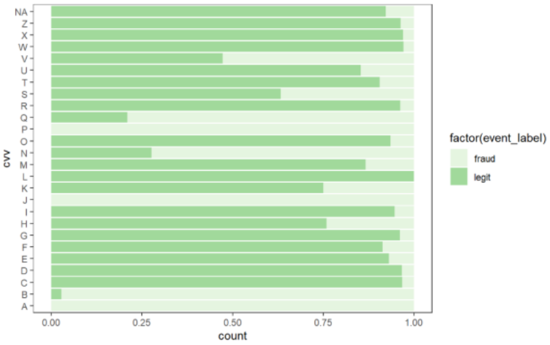


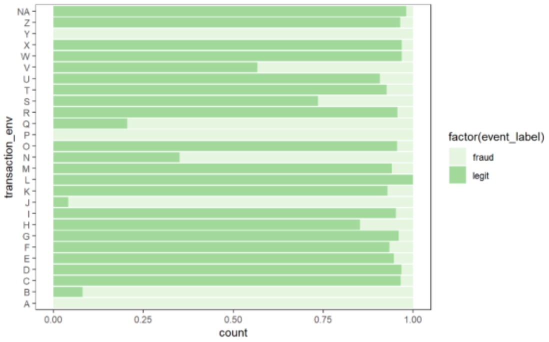
By box plotting all numerical variables, several less frequent numerical variables were identified that would be considered categorical. In addition, boxplots compare the distributions of fraud and legit, and some boxplots show comparative differences, implying that these characteristics of customers may be associated with encountering fraud. For example, the overall distribution of account opening times for customers who experienced fraud was higher than for customers who were not fraudulent (lower mean and quartile).In a business environment, this could mean that older accounts are more vulnerable to fraud, possibly because such customers are more likely to have neglected account management.

Also, similarly, there are differences in the distribution of individual transfer amounts. The larger the amount of a single transfer, the easier it is to be defrauded, possibly because the cost of defrauding a small-amount transfer is higher than the benefit. Furthermore, the similar shapes of these two variables may indicate a high correlation between them, so multicollinearity should be considered in later regression models.

3.3. Exploring categorical variables







A 100% stacked column chart is used to explore differences in scams based on categorical variables. For example, transactions of different currencies have a different probability of being defrauded. Different cvv codes also have different chances of being scammed. Different transfer environments also have different possibilities for fraud.

4. Methodology

4.1. Preparing data

4.1.1. Making factors
4.1.2. Partition data to 70/30

4.2. Decision Tree

4.2.1. Define the recipe for DT
4.2.2. Define the model, create a workflow, and fit the model
4.2.3. Scoring training and testing datasets

4.3. Random forest 1

4.3.1. Define the recipe for RF
4.3.2. Define the model, create a workflow, and fit the model
4.3.3. Scoring training and testing datasets

4.4. Random forest with email

4.4.1. Define the recipe for RF
4.4.2. Define the model, create a workflow, and fit the model
4.4.3. Scoring training and testing datasets

5. Model Metrics & Evaluation

The following metrics will be used for evaluation:

- **Accuracy**
  Accuracy is simply a measure of how often the classifier correctly predicted. It can be defined as the ratio of the number of correct predictions to the total number of predictions.

$$\textbf{\textit{Accuracy}} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **ROC_AUC**
  The Receiver Operator Characteristic (ROC), which separates the "signal" from the "noise," is a probability curve that compares the TPR (True Positive Rate) against the FPR (False Positive Rate) at different threshold values.
  A classifier's capacity to distinguish between classes is measured by the area under the curve (AUC).
- **Precision**
  Precision describes the proportion of accurately predicted cases that actually resulted in a positive outcome. Precision is helpful when False Positives are more problematic than False Negatives.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

- **Recall**

  Recall describes how many of the actual positive cases our model was able to accurately anticipate. When a False Negative is more important than a False Positive, it is a valuable metric.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

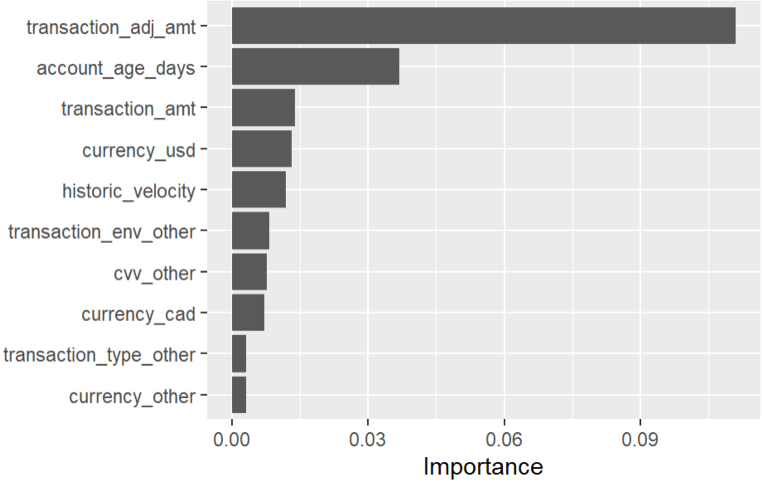
5.1. Evaluating

Model	Partition	AUC	Precision	Recall
Model 1-rf	train	0.9576816	0.6007067	0.7549165
Model 2 - rf_email	train	0.9562933	0.6001338	0.7585113
Model 3 -dt	train	0.8786074	0.4541695	0.6967646

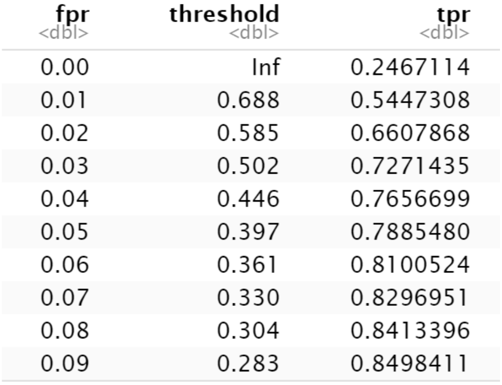
Model	Partition	AUC	Precision	Recall
Model 1-rf	test	0.9313354	0.5820085	0.7300584
Model 2 - rf_email	test	0.9303860	0.5804088	0.7320039
Model 3 -dt	test	0.8748908	0.4469291	0.6901751

Based on various evaluation metrics, we can assume that the accuracy of random forest is higher than that of decision tree model.

5.2. Feature Importance by Model



5.3. Selected Model Operating Ranges

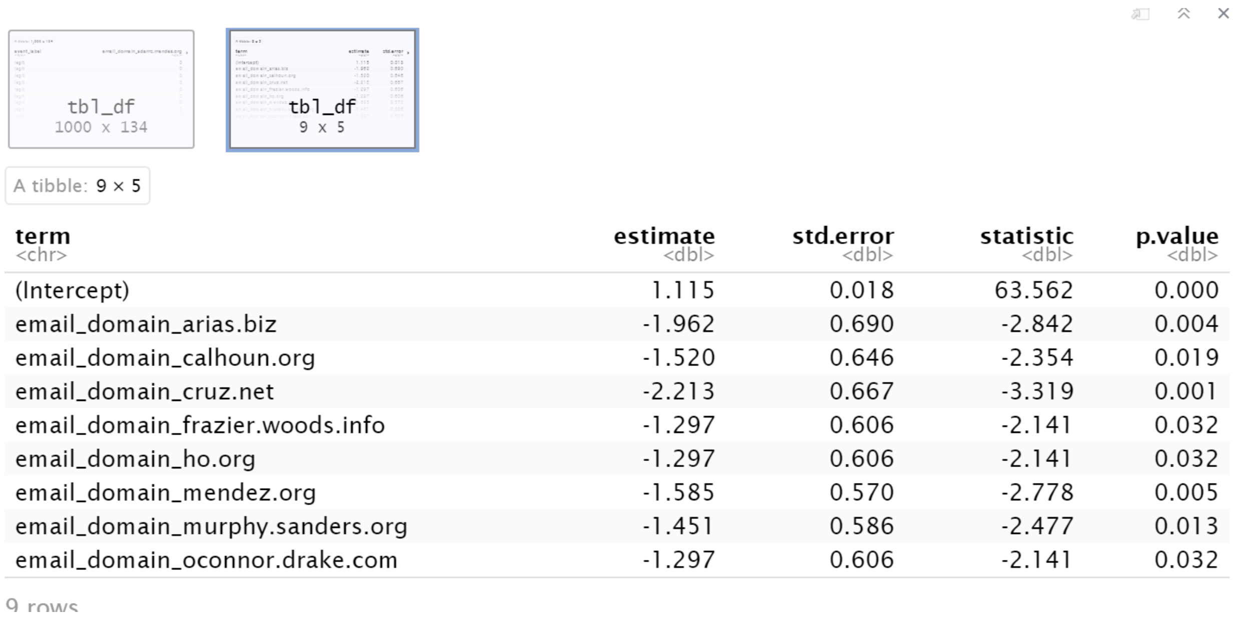


fpr <dbl>	threshold <dbl>	tpr <dbl>
0.00	Inf	0.2467114
0.01	0.688	0.5447308
0.02	0.585	0.6607868
0.03	0.502	0.7271435
0.04	0.446	0.7656699
0.05	0.397	0.7885480
0.06	0.361	0.8100524
0.07	0.330	0.8296951
0.08	0.304	0.8413396
0.09	0.283	0.8498411

6. Discussion & Recommendations

6.1. Discussion about email_domain and billing_postal

6.1.1. The Firm believes that email domain and billing postal code is an important predictor, your write up should discuss why or why not.

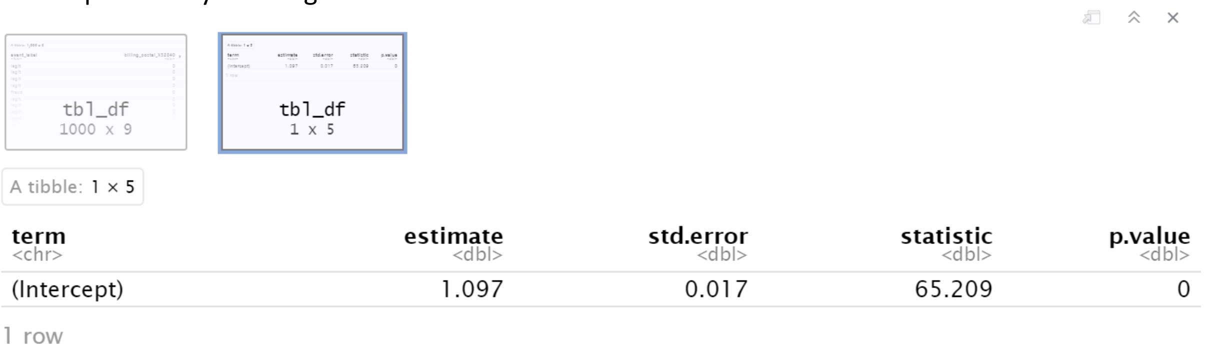


A tibble: 9 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	1.115	0.018	63.562	0.000
email_domain_arias.biz	-1.962	0.690	-2.842	0.004
email_domain_calhoun.org	-1.520	0.646	-2.354	0.019
email_domain_cruz.net	-2.213	0.667	-3.319	0.001
email_domain_frazier.woods.info	-1.297	0.606	-2.141	0.032
email_domain_ho.org	-1.297	0.606	-2.141	0.032
email_domain_mendez.org	-1.585	0.570	-2.778	0.005
email_domain_murphy.sanders.org	-1.451	0.586	-2.477	0.013
email_domain_oconnor.drake.com	-1.297	0.606	-2.141	0.032

9 rows

By performing logistic regression on email domain with the target variable, we can find that 5 out of 134 levels are significant, at a level of significance of 0.05. And when the more common email domain is added to the random forest model for operation, we find that the accuracy and roc-AUC have decreased. From these phenomena, we have reason to believe that email domain has nothing to do with the probability of being scammed.



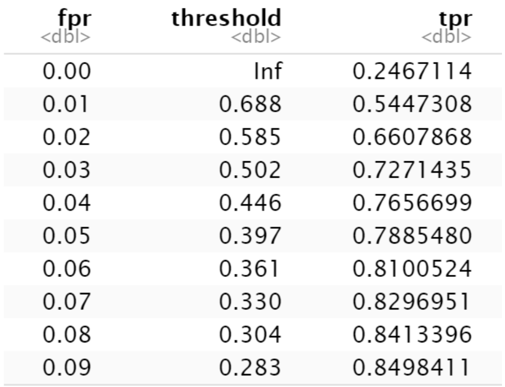
A tibble: 1 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	1.097	0.017	65.209	0

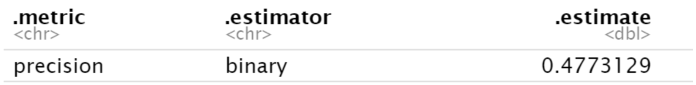
1 row

In the same way, performing logistic regression on billing_postal and the target variable, we can find that there are no significant levels in the 9 levels, at a significant level of 0.05. We can infer that it has nothing to do with the probability of encountering a scam.

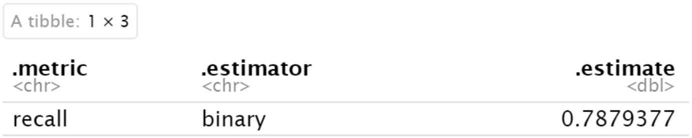
6.1.2. Finally, the firm wants to operate at a 5% false positive rate, based on your best-performing model what is the rule a with score threshold that will give them a 5% false positive rate, what is the recall and precision at that threshold?



fpr <dbl>	threshold <dbl>	tpr <dbl>
0.00	Inf	0.2467114
0.01	0.688	0.5447308
0.02	0.585	0.6607868
0.03	0.502	0.7271435
0.04	0.446	0.7656699
0.05	0.397	0.7885480
0.06	0.361	0.8100524
0.07	0.330	0.8296951
0.08	0.304	0.8413396
0.09	0.283	0.8498411



.metric <chr>	.estimator <chr>	.estimate <dbl>
precision	binary	0.4773129



A tibble: 1 × 3

.metric <chr>	.estimator <chr>	.estimate <dbl>
recall	binary	0.7879377

When the FPR is equal to 5%, the threshold is equal to 0.397 and the TPR is equal to 0.789. Meanwhile, precision and recall are equal to 0.477 and 0.788.

5% FPR means that there is a 5% probability that the model thinks that a transaction may be defrauded, but it does not happen. At the same time, the probability of correctly predicting a fraudulent transaction is 78.9, based on my best model.

Based on my model, my precision is below 50%, which means the model's ability to predict fraud is somewhat poor. But the recall is higher than 0.7, which means that the model has identified 70% of the frauds.

6.2. Recommendations

- The company has a high degree of attention to the two variables email domain and billing postal, but these two variables have been statistically verified, and we found that they may not be related to encountering fraud. So, when trying to predict scams, we should not take these two variables too much into consideration.
- Because as mentioned in the exploratory analysis, the longer the account is opened, the more likely it will be fraudulent, so I suggest that the bank should remind them to pay attention to the safety of the transfer by email or phone for customers who have opened the account for a longer time.
- Transfers with a large single transfer amount are also more prone to fraud, so for transfers exceeding a certain limit, the bank should use multiple verification methods to ensure that the transfer is made by myself and that the transfer occurs in a safe condition.
- Because specific transfer scenarios also seem to be related to the possibility of encountering fraud, we recommend that companies implement security measures for transfer environments with high fraud encounters to ensure greater transfer security.

7. Kaggle Submission

Kaggle Name: Charlene Wei

Kaggle reported score: 0.92286

Kaggle reported position at time of submission: #116

(Note: this will change as others post)

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