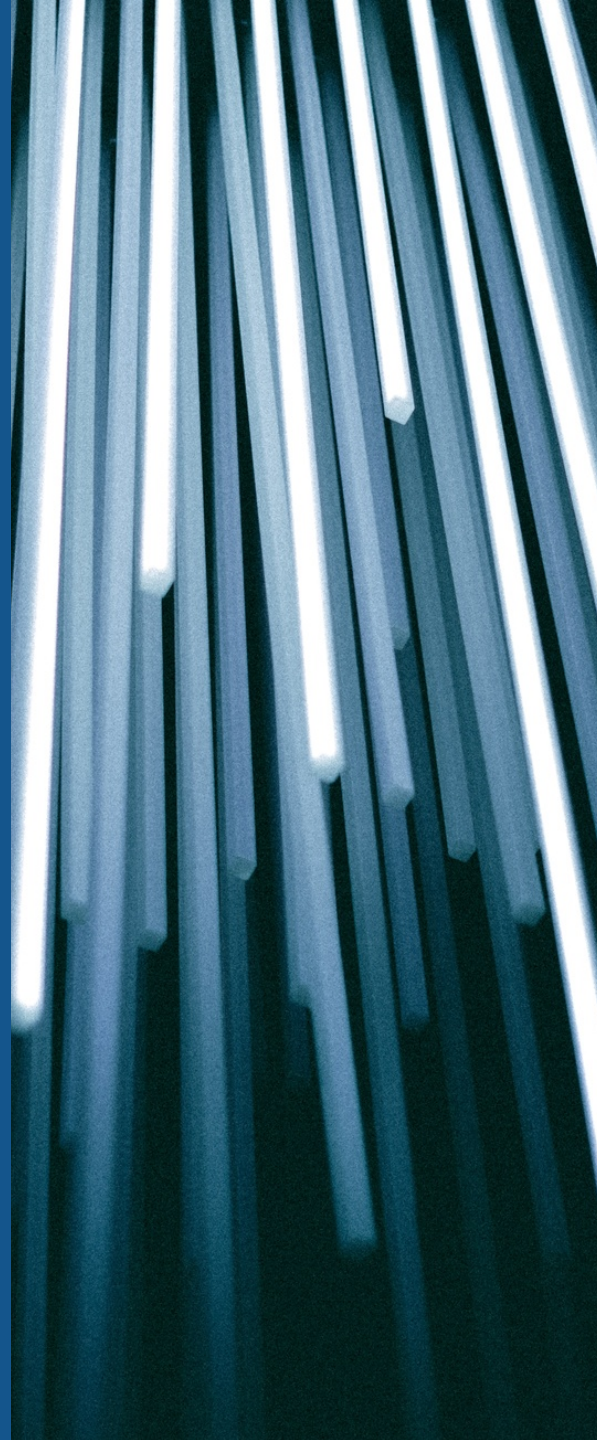
CAPITAL ONE DATA CHALLENGE

Card Transaction

Summary



Overview

Frist

- Enhance features by Feature Engineering.
- Identify important features using machine learning.

Second

- Insights and visualization

Last

- Modeling
- Next Steps

Data Exploratory Analysis

We have **786363** rows and **29** columns

Data is from **5000** accounts

62.6% of fraud loss happened in top 20 merchant

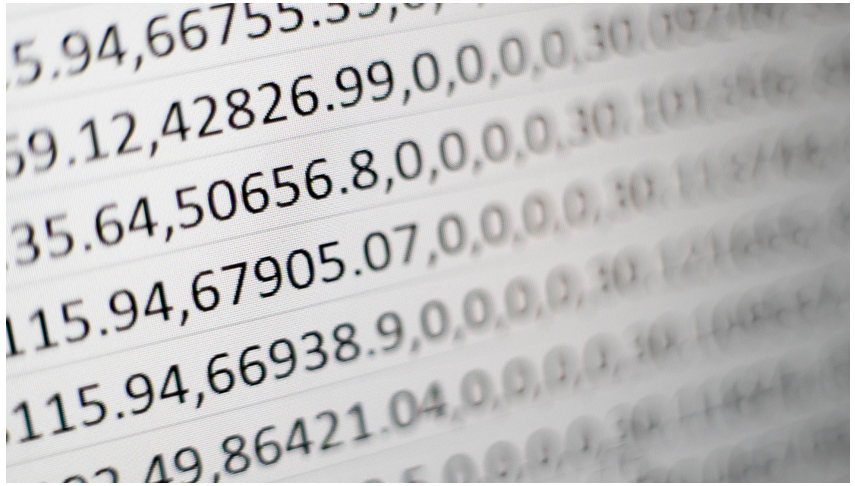
Over **99%** of transaction happened in US

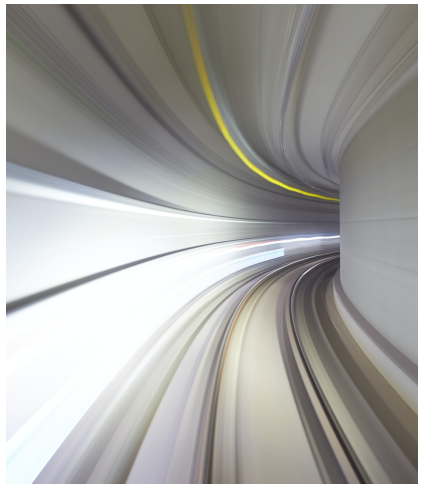
Around **1.5%** of transaction is fraud

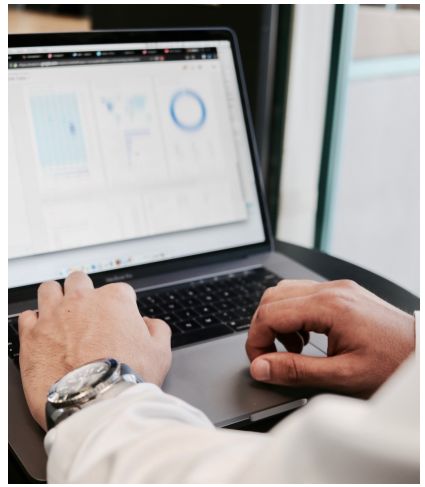
The average transaction amount between Reversal Transaction and Multi-transaction is similar

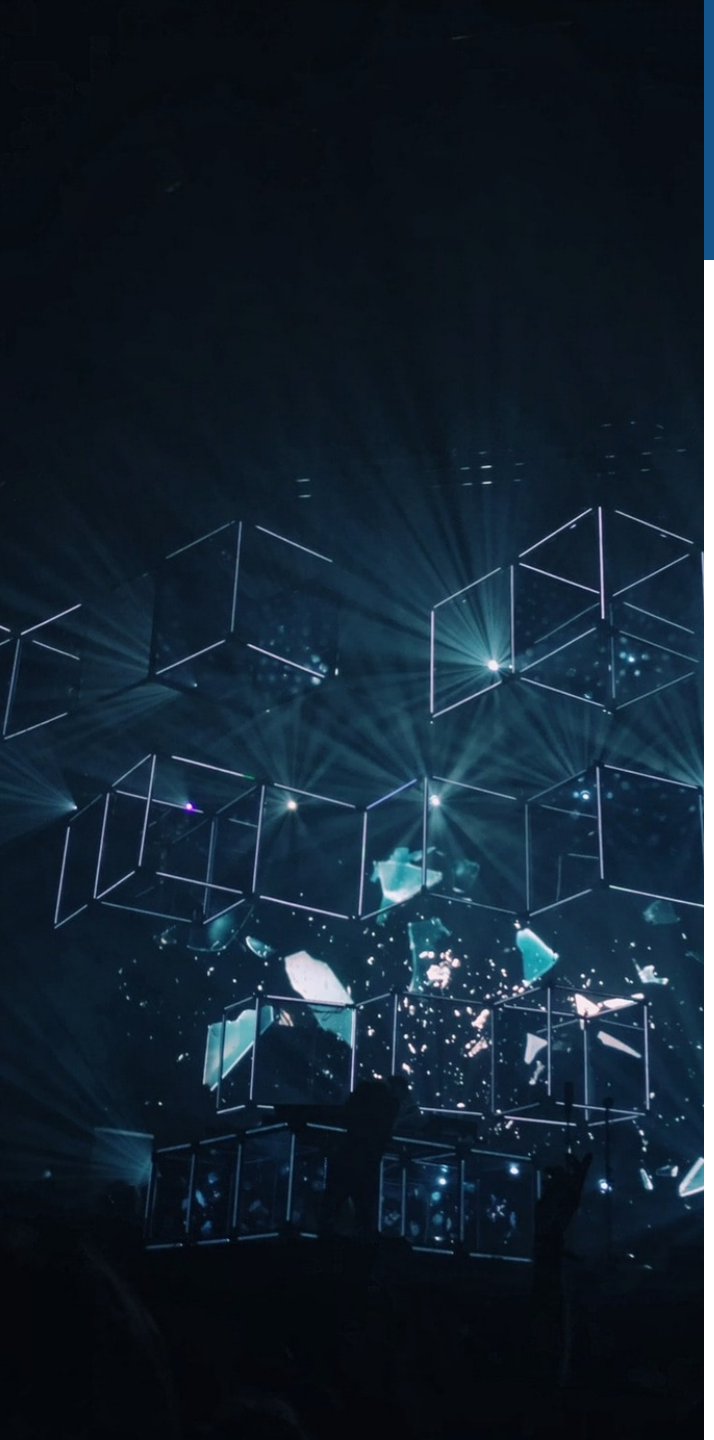
Data Processing

- Reversal transaction flag
- Muti-transaction flag
- Outside United State transaction flag
- Days from open account to transaction date
- Percentage of transaction amount in available amount
- Percentage of transaction amount in credit limit









Feature Selection

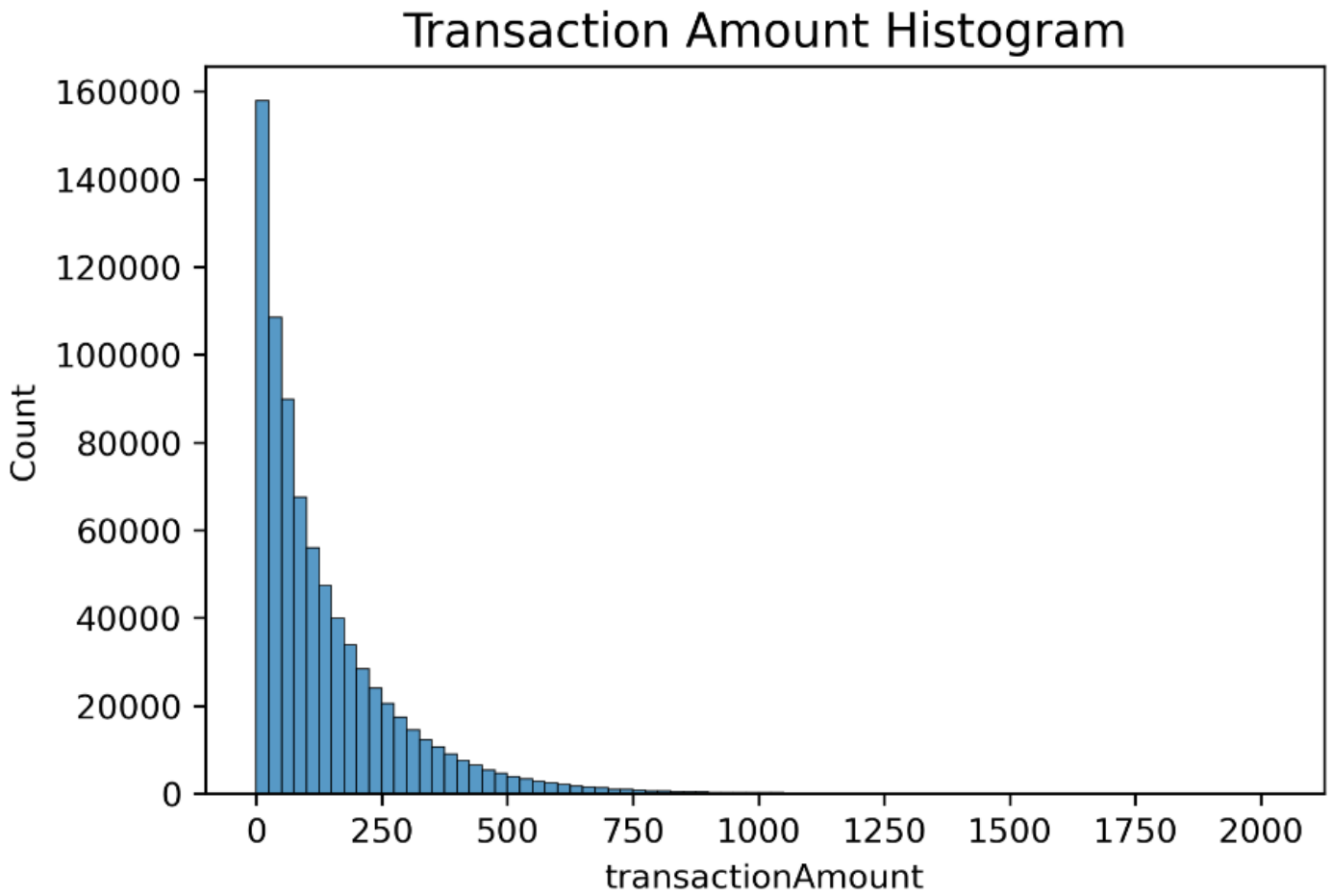
Goal: To identify significant factors that affect cancellation behaviors of riders

Methodology: Logistic regression, Random Forest, & Xgboost model

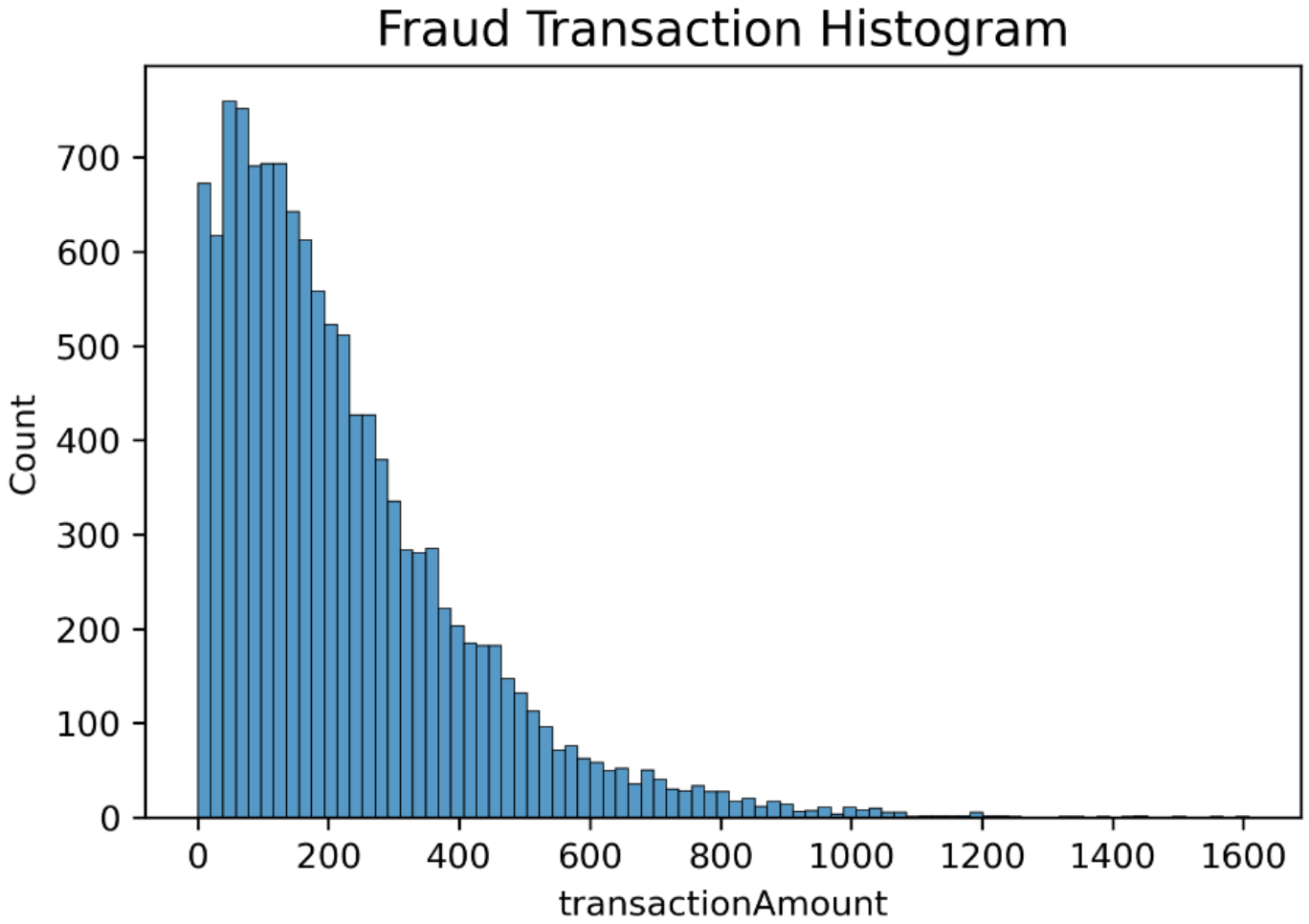
Factors included in both models:

- creditLimit
- availableMoney
- currentBalance
- cardPresent
- expirationDateKeyInMatch
- Duplicated
- is_reversal_duplicate
- is_multi_transaction
- date_diff_trans_open
- outside_US
- percentage_trans_available
- percentage_trans_limit

Plots


Transaction Amount Histogram
160000
140000
120000
100000
80000
60000
40000
20000
0
Count
0
250
500
750
1000
1250
1500
1750
2000
transactionAmount

Plots


Fraud Transaction Histogram
Count
0
100
200
300
400
500
600
700
0
200
400
600
800
1000
1200
1400
1600
transactionAmount

Finds

- The significant difference between the histogram of the fraud plot and the transaction plot happens in large amount transaction(**over 200$**).
- We can assume that the transaction amount is a important feature to determent whether a transaction is fraud or not.

Models

- ## XG-Boost

  XG-Boost can automatically weight data, which means it can work well in the imbalanced dataset (fraud detection).

- ## Random Forest

  Random Forest is good at handle high-dimensional datasets with a large number of features(we have xx features in this dataset). But for the imbalance data we need to resample the data first.

- ## Logistic Regression

  Logistic regression is a statistical algorithm for binary classification, which is also workable in fraud detection. Like random forest model, we need to resample the imbalance data first.

Model Performance

Logistic regression

Mean ROC AUC: 0.500

Accuracy Score: 97.326

Random Forest

Mean ROC AUC: 0.625

Accuracy Score: 97.326

Tree-Based XG-Boost

Mean ROC AUC: 0.707

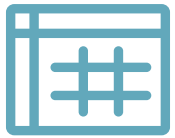
Accuracy Score: 98.445

Top 3 features

3 most important features

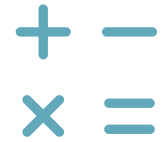
Card Present

When card present, there are less probability that this transaction is a fraud



Transaction Amount

Larger transaction amount have more probability to have fraud



Percentage of Transactions within Credit Limit

When this percentage is high, the transaction have more probability to be fraud

Questions

- Why the average transaction amount between the **multi-transaction(146.7$)** and **reversal transaction(147$)** is similiar?
- Is there any **overlap** between them in the real world?
- These days, more and more people use their credit card online, beside **card present,** is there any other **potential features** can be as reliable as **card present**?

1 Test the distributions between Fraud amount and all transaction

2 Make a potential fraud flag for users who have been fraud more than once.

3 Add geographic feature into model

4 Add online and on-store flag

5 Make a model only for the transaction over 200$

Next Steps

Thanks for Watching