Fraud Detection

Data

- Credit card transactions from a credit union.
- Initial size: 41 million, covers 2000 distinct cards. Transaction 01/2018 to 12/2019
- Resample: full history of 100 cards, 120K instances
- Variables:
  - Merchant: address and type of business
  - Customer: address of residence, basic demographic data, and employment
  - Transaction: amount, time, and credit card number
- Imbalanced class: less than 1% is positive
  - Up-sample to 20%, to reduce randomness caused from data points when splitting train/val

Feature

- Customer level:
  - Card holder's age
  - Average income from employment information
- Transaction:
  - Distance between merchant address and billing address
  - Time of day when transaction happened (morning, afternoon, evening, late night)
- Time-series:
  - Number of transactions of same card from last 1 hour, 24 hours
  - How many std from the average of personal purchase history

Model

- Baseline: random guessing
- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting

Evaluation

- AUC, F-1 score
- Optimizes recall in order to minimize the false negative rate
- Output:
  - Decision Tree: 81%
  - Random Forest: 87.5%
  - Gradient Boosting: 89.2%
  - Favor random forest since it's faster

Further Analysis:

- Most important features: transaction amount, distance, number of transactions in last 1 hour
- Learning curve: to observe whether sample size is good enough
  - Curve itself is ok
  - But due to up-sampling strategy, give more time and computing resources, would sample more cards (especially positive cases), and down-sample negative cases.