Fraud Detection: A Topic Ever More Important in the COVID-19 Era

—

Shih-Kai Lin

May 24, 2021

The Data Incubator Interview

Motivation: e-commerce is on the rise

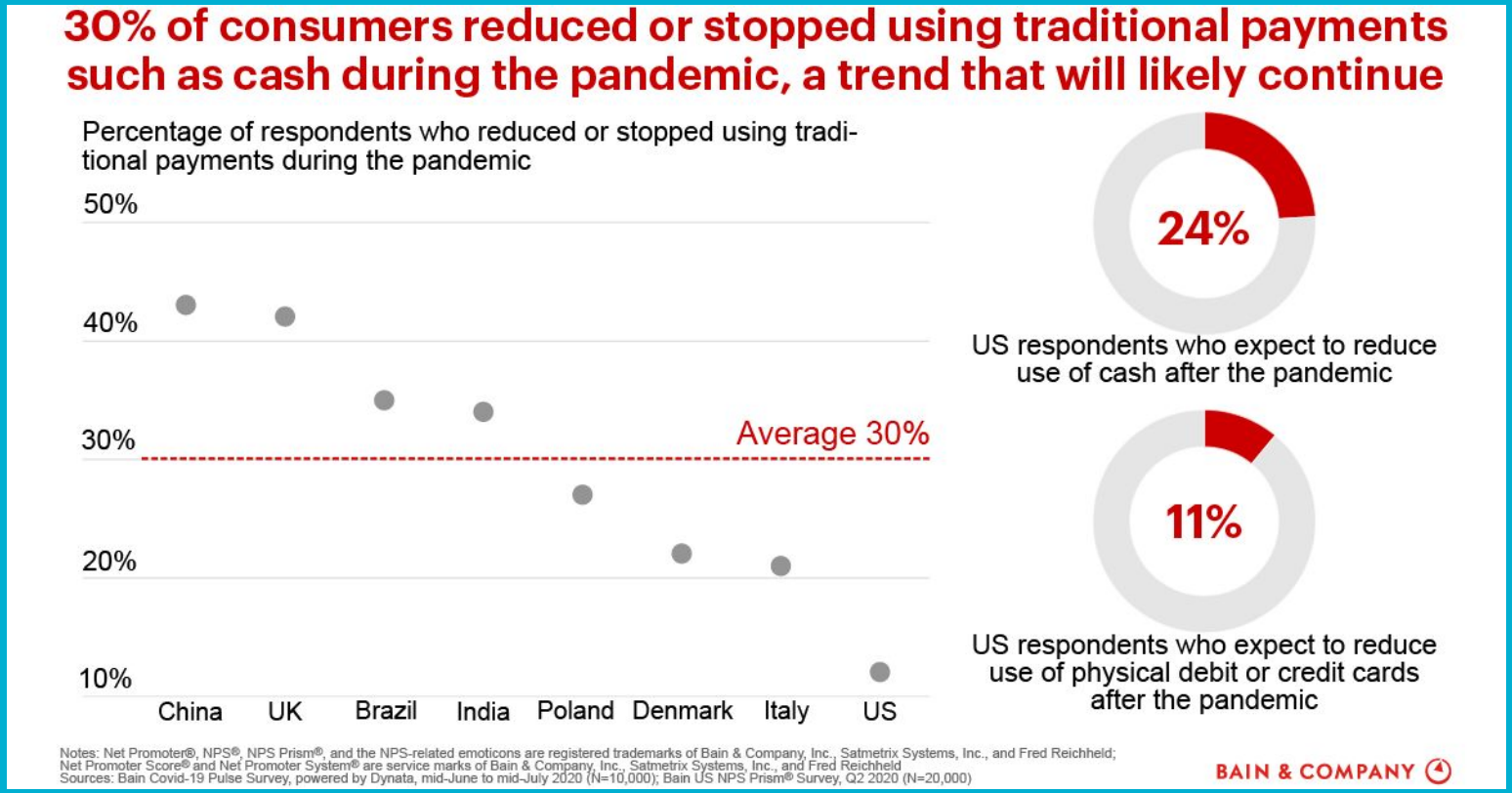
Source: Brookings

Figure 1. Worldwide Google searches and COVID-19

Source: Google Trends topic interest over time (normalized to index with scale 0-100); Johns Hopkins Coronavirus Resource Center

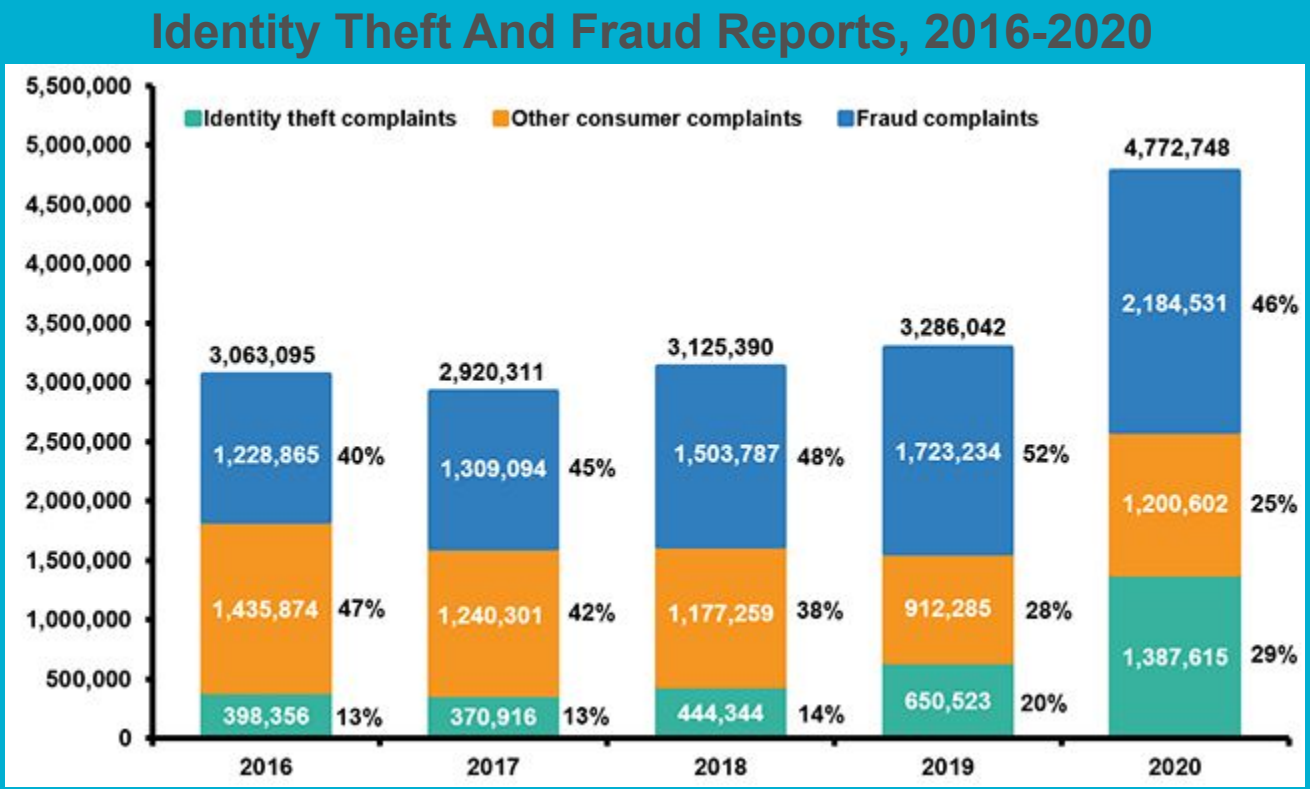
BROOKINGS

COVID-19 will permanently shift online shopping habits



Source:
[Bain & Company]

Fraud is on the rise during the pandemic

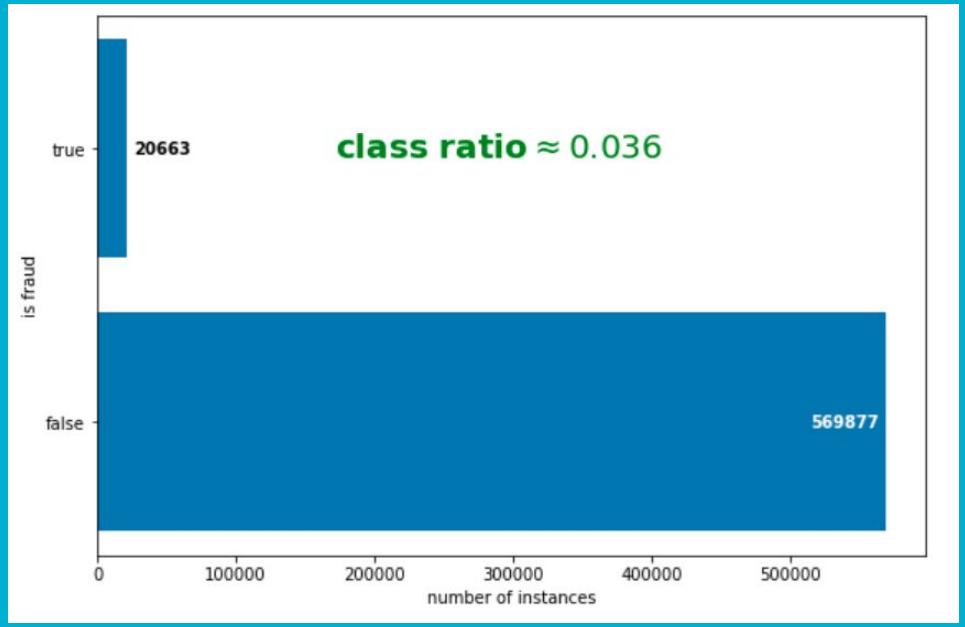


Source: [Insurance information institute]

Project proposal & dataset: Detection of online fraudulent transactions

- Dataset: IEEE-CIS Fraud Detection
  - The data comes from Vesta's real-world e-commerce transactions and contains a wide range of features from device type to product features.
- Available on Kaggle.
  - API command:
    ```
    $ kaggle competitions download -c ieee-fraud-detection
    ```
- Files to be used for this project
  - train_transaction.csv - the transaction table
  - Train_identity.csv - the identity table
- Goal: Given variables for a transaction, develop a machine learning model to predict whether an online transaction is genuine or fraudulent.

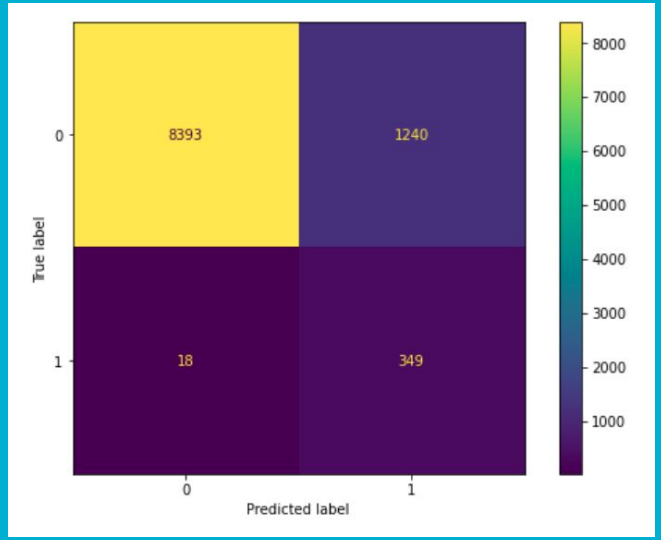
Imbalance classification: a hard problem



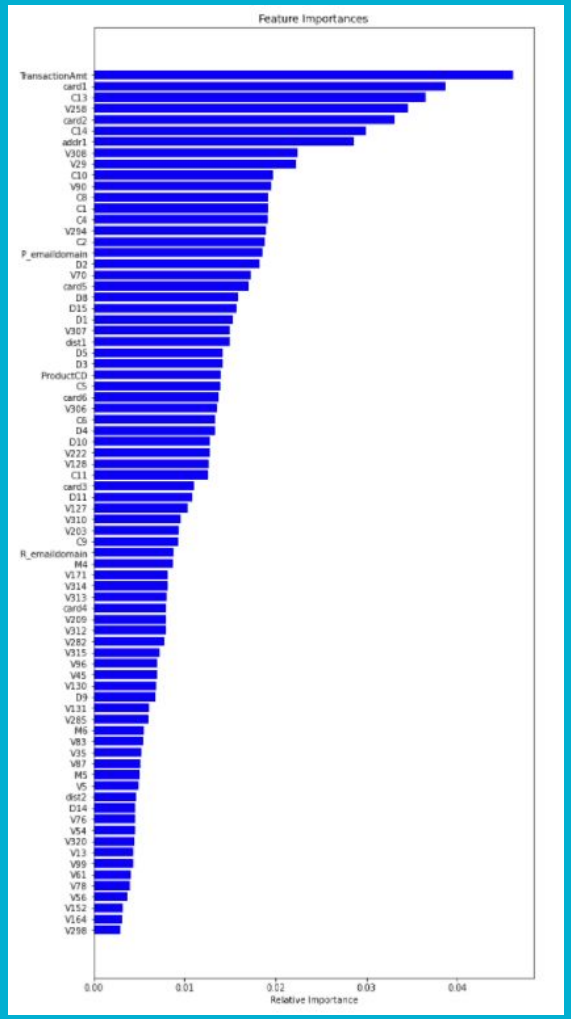
- Fraudulent transactions are much fewer than normal transactions - a typical case for imbalanced classification.
- Where the number of instances of one class is much fewer than the other.
- Hard problem for machine learning algorithms: simply assigning all instances to non-fraud results in a fantastic model of 96.4% accuracy - no detection at all!
- Many methods are developed to address imbalance classification. This project will look for one that performs as well as possible.

A first attempt with random forest & undersampling

- The famous machine learning library, `scikit-learn` is used
- The performant ensemble method, random forest is used, with decision tree as the base classifier
- True positive rate HIGH 👍
  True negative rate HIGH 👍
  False positive rate HIGH 👎
- A byproduct is the feature importance ranking



TP: 349/(349+18)=95%
TN: 8393/(8393+1240)=87%
FP: 1240/(1240+349)=78%



Future Work

Of course, there is a lot of room for improvement. Below is a list already on my mind.

- Use information in the identity table.
- Investigate other undersampling techniques.
  - They might not be feasible anyway. Afterall, I have been running the Condensed Nearest Neighbor algorithm for days offline without any output.
- Try oversampling techniques (eg. Synthetic Minority Oversampling TEchnique)
- Investigate other dimensionality reduction technique, such as autoencoders.
- Utilize more genuine transactions for model training.
  - One idea I have now is to partition the genuine transactions into groups of the same size as the fraudulent transactions. Train one classifier for each group, and combine the models with ensemble methods. This procedure might be equivalent to oversampling the minority class.
- Try other classifiers or implementations, such as XGBoost.