

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

Shih-Kai Lin

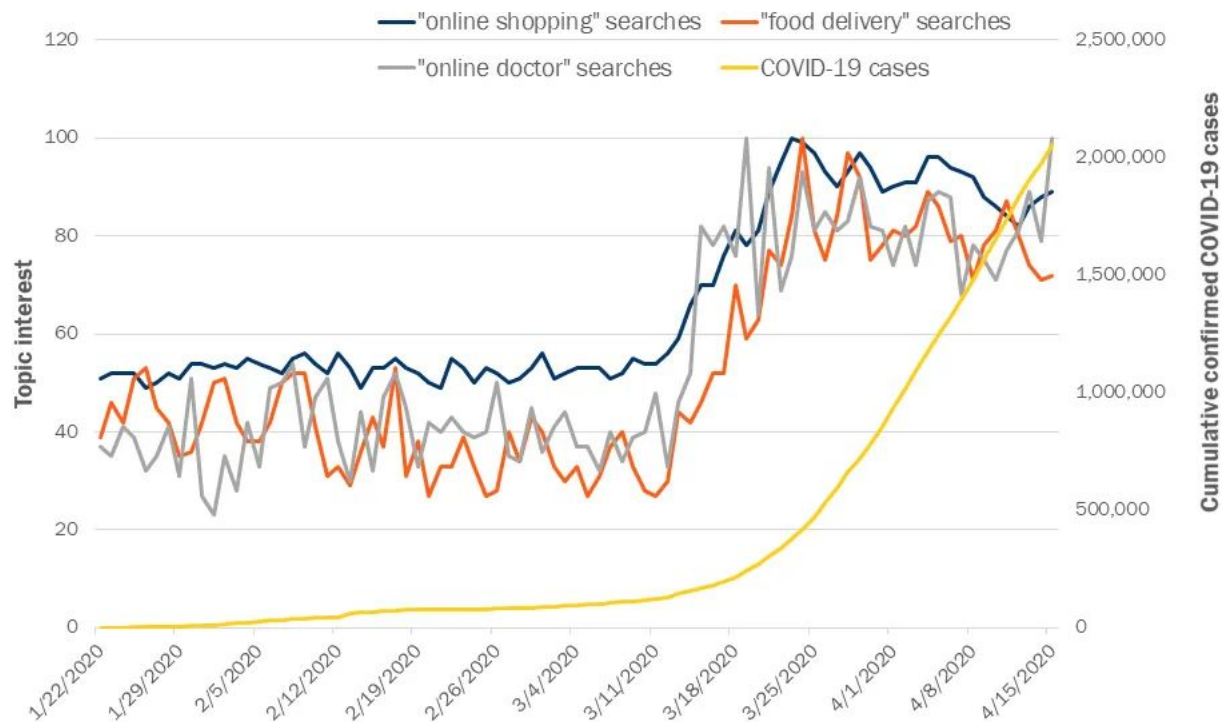
May 24, 2021

The Data Incubator Interview

Motivation: e-commerce is on the rise

Source: [Brookings](#)

Figure 1. Worldwide Google searches and COVID-19



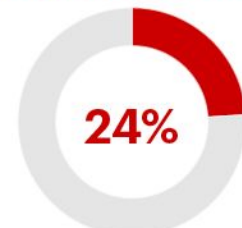
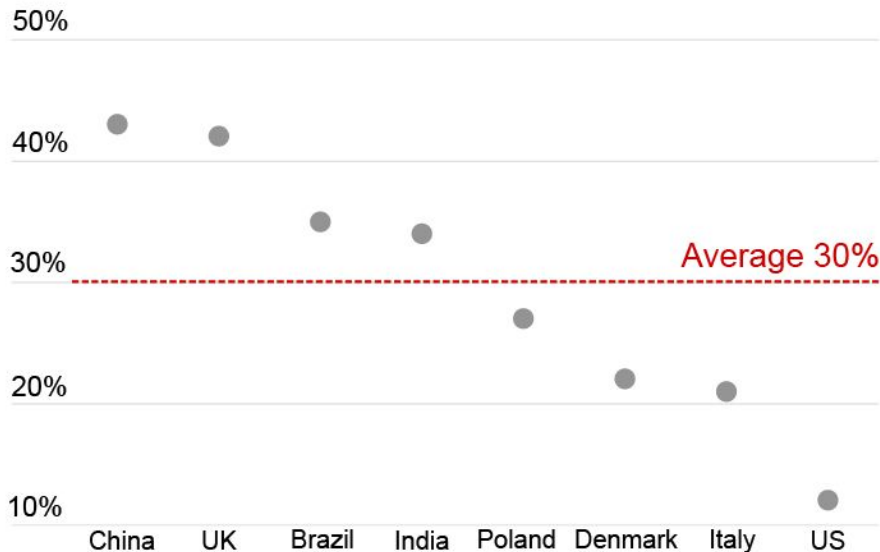
Source: Google Trends topic interest over time (normalized to index with scale 0-100);
Johns Hopkins Coronavirus Resource Center

BROOKINGS

COVID-19 will permanently shift online shopping habits

30% of consumers reduced or stopped using traditional payments such as cash during the pandemic, a trend that will likely continue

Percentage of respondents who reduced or stopped using traditional payments during the pandemic



US respondents who expect to reduce use of cash after the pandemic



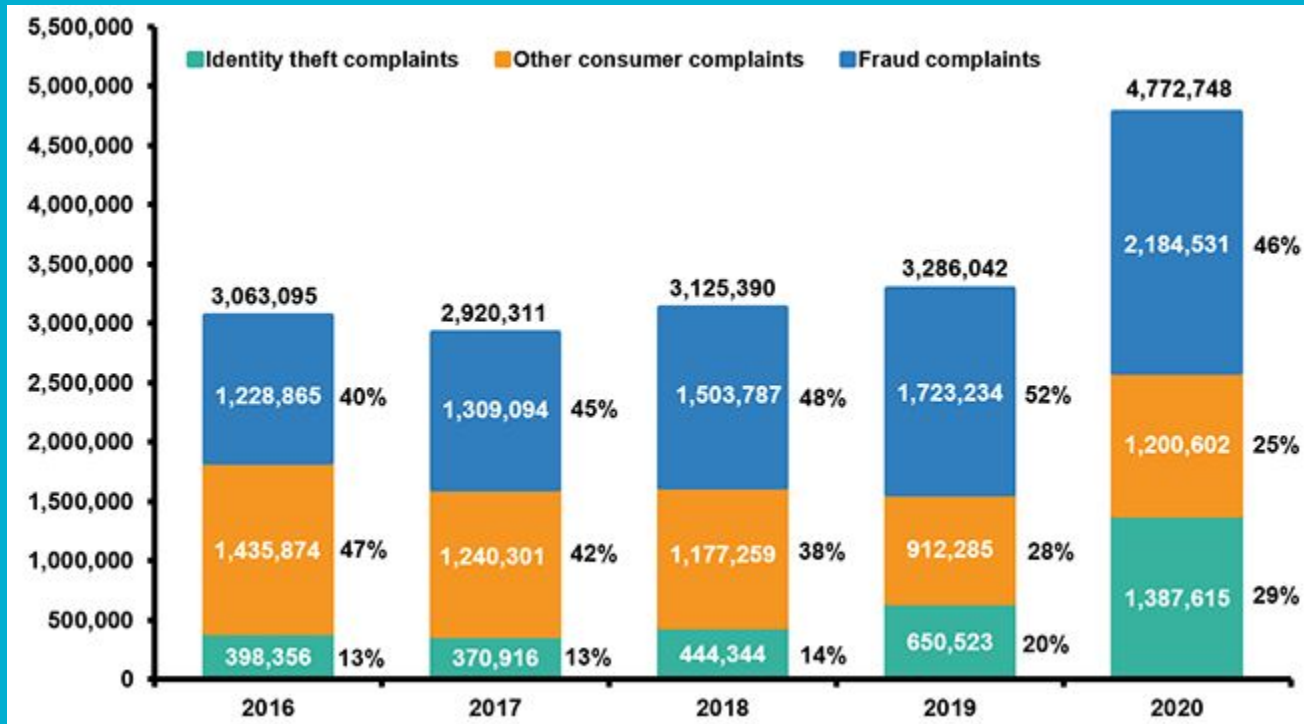
US respondents who expect to reduce use of physical debit or credit cards after the pandemic

Source:
[Bain & Company](#)

Notes: Net Promoter®, NPS®, NPS Prism®, and the NPS-related emoticons are registered trademarks of Bain & Company, Inc., Satmetrix Systems, Inc., and Fred Reichheld; Net Promoter Score® and Net Promoter System® are service marks of Bain & Company, Inc., Satmetrix Systems, Inc., and Fred Reichheld
Sources: Bain Covid-19 Pulse Survey, powered by Dynata, mid-June to mid-July 2020 (N=10,000); Bain US NPS Prism® Survey, Q2 2020 (N=20,000)

Fraud is on the rise during the pandemic

Identity Theft And Fraud Reports, 2016-2020



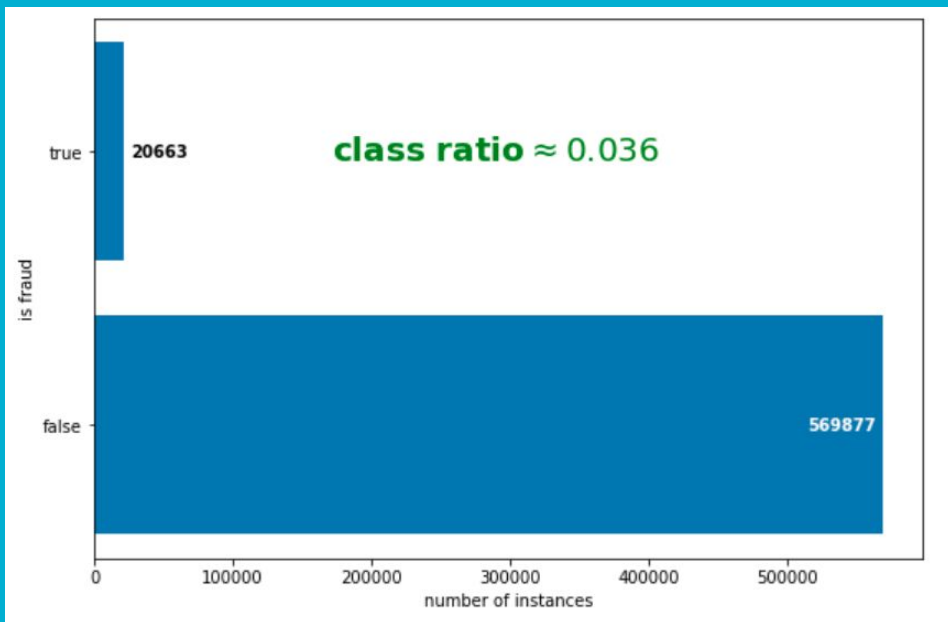
Source:
[Insurance
information
institute](#)

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.

Imbalanced 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.

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.