Capstone 2 Project Proposal: Identifying Credit Card Fraud

Mark Cohen

Problem Statement

What features of a credit card transaction can be used as indicators that it is likely to be fraudulent?

Context

It is estimated that fraud involving credit and debit cards causes losses of nearly $30 billion worldwide, as of 2019.[1] To combat this, card issues and networks need to be proactive in identifying and stopping fraudulent transactions in real time.

Criteria for Success

Produce a model that can:

1. assess the risk of fraudulence of individual transactions based on features of the transaction itself and customers' transaction history, and
2. flag cards to be locked for customer verification.

Scope of Solution Space

Because of the importance of real-time fraud identification, the final model needs to be entirely backwards looking, relying on only the characteristics of each transaction and those that came before.

Constraints

False positives represent a nuisance to legitimate card-users. The benefit of catching a larger share of fraudulent transactions needs to be weighed against this cost.

Stakeholders

This problem is likely to arise in the context of a consumer bank or credit card company. The key internal stakeholders would likely include:

- Security teams
- Legal and compliance teams
- Customer service teams

Data Sources

At team at IBM simulated over 20 million transactions by 2,000 U.S.-based customers over

[1] https://www.cnbc.com/2021/01/27/credit-card-fraud-is-on-the-rise-due-to-covid-pandemic.html

multiple decades.[2] The advantage of synthetic data is that it can include information that would risk identifiability in real-world data, which thus cannot be publicly shared due to privacy concerns.

Methods

It would be premature to determine the details of the modeling strategy, but the two preliminary points can be made:

1. Since 20 million transactions would not be tractable on the limited computing resources available, the model will be trained and tested on a subsample.
2. Although this is, ultimately, a classification problem, it is desirable to think not in terms of binary estimate of "fraud or not" but instead in terms of probabilities, i.e. "risk of fraudulence."

Key Deliverables

This project will deliver three products:

1. The model itself, encapsulated in notebooks available on a github repository.
2. A report describing the model and outlining the findings of the modeling process.
3. A high-level slide deck for presenting the main upshots to non-technical stakeholders.

[2] Erik R. Altman. 2019. "Synthesizing Credit Card Transactions." https://arxiv.org/abs/1910.03033