Benjamin Katz

Week 5

Oct 16, 2023

Colab: <https://colab.research.google.com/drive/1jUSF4vwt6BLiilRjfTZyYGWmtCYXkvEs?usp=sharing>

The most important part of fraud detection is detecting the fraud before allowing the transaction to take place. With that in mind, the model must be productionizing such that any card reader can make a call to the model and get instant authorization. This would be realtime. In theory, this model can be hosted at the POS system but since the POS needs to make a call to the credit card company for other authorization purposes(credit limit) it makes sense to host the fraud detection model in the same place. The fraud detection service can be hosted on a traditional server or as a lambda on AWS or other similar cloud based services.

Tensor flow extended is an ML pipeline that allows for the importing and cleaning of data as well as the training for models. It has built in scalability to serve application endpoints when the need grows. Apache Kafka and Flink have the advantage of being open source and utilizes kafka which is optimized for data streaming which would work well for a real time service like the fraud detection system.

Retailers can call this service with standard http calls to quickly determine if the transaction is fraudulent. The issue of scaling is often solved by cloud based services but it is an easily parallelizable problem. If many clients are interacting with the model at a given time, simply add more servers with the model on them. Every time fraud is detected there needs to be a way for the user to either cancel the credit card or report a false positive. To do this, a user could receive a text message and reply with either a fraud or no fraud. This helps the now verified user continue with her purchase as well as provide a ground truth for monitoring metrics of the model. With these ground truths we can monitor precision and recall a low precision would mean that the customer is constantly getting her purchases rejected which would be annoying for her. A low recall would mean the credit card company loses money to fraud.

Because scammers are constantly evolving, the model for catching them needs to evolve as well. Utilizing the ML pipeline, the service can update the model in batches and alert if any metrics start to shift over time to see if the model architecture needs to be changed as new scam methods are developed.