Bio

György Móra is principal data scientist at Ekata (formerly Whitepages Pro) where he works on the machine learning solutions powering Ekata's Identity Check Confidence Score and Transaction Risk Score. He has a background in software engineering and natural language processing research. His main interest in ML is how to deliver value to users.

Abstract

At Ekata, we evaluate identity information in online shopping transactions to enable our customers (vendors) to protect against fraud. We make commitments to deliver pre-purchase verification predictions with very low latency.

When we built the ML system powering the Identity Check Confidence Score, we faced the following challenges:

- The pipeline that trains our model had to digest a large amount of transaction data
- We had to perform a large set of validation experiments to ensure each new release caused no disruptions for our customers
- At the same time, the model had to perform in a real time, low latency environment

In this talk, I discuss how we met these challenges by building a Spark- and XGBoost-based training and experimentation pipeline that delivers our models as a custom predictor library, to allow seamless integration with the production environment. Our embeddable library enables super-fast predictions real-time.

There are existing formats for exporting and storing ML models in order to load them into the production system to deliver predictions. But in these existing formats feature extraction and normalization usually need to be reimplemented in the production systems, and are not part of the model description. Our solution encapsulates feature extraction, normalization, and prediction all in one unit. This gives the data science team a lot more flexibility to make changes and permits easy integration for engineering teams.