Cecilia Bell 6.3.2021

Classification Project Proposal

Predicting KKBOX User Churn

Question

For subscription businesses, reducing churn can drastically improve revenue. Accurately predicting if a user will churn, then, serves as a key step towards increasing customer retention rates and long-term business success. For this project, I plan to build a binary classification model to predict whether KKBOX users will churn. KKBOX is Asia's leading music streaming service and holds the world's largest Asia-Pop library. KKBOX offers users unlimited access to its vast library based on both advertisements and subscriptions, but the business model relies on retaining paid users.

Data

I will use data supplied by KKBOX for the 11th ACM International Conference on Web Search and Data Mining (WSDM 2018), which can be downloaded from Kaggle. 10 datasets are available:

- two containing data on whether a user churned
- two containing data on each user's transactions (payment method, payment plan days, etc.)
- two containing data on each user's daily behaviour (number of unique songs played, etc.)
- two containing data on each user (city, age, etc.)
- two sample submissions

The first set in each of the first four pairs contains data from 2016 up to March 2017, while the second just contains data from March 2017 (the second sets were created after the original March test sets were leaked). Given the large size of the datasets, which total over 8GB, I will only use the March 2017 datasets as both training and test data. The key features I expect to work with include transaction date, membership expiration date, and is_cancel. Notably, cancelling a subscription does not imply that a user has churned; a user is only considered to have churned when they do not start any new subscription within 30 days after cancelling their last subscription.

Tools

I plan to use Python packages such as Scikit-Learn and XGBoost to build various classification models, which I will then cross-validate before choosing the best model. I will also use Tableau to visualize unexpected insights.

Cecilia Bell 6.3.2021

MVP Goal

My MVP goal is to build one classification model, such as a logistic regression model, using a few features. In order to build the model, I will need to have cleaned and aggregated the data, and performed some preliminary exploratory data analysis.