KKBOX Subscription Service Churn Prediction

Context

KKBOX is Taiwan's leading music streaming service, holding the world's most comprehensive Asia-Pop music library with over 30 million tracks. They offer a generous, unlimited version of their service to millions of people, supported by advertising and paid subscriptions. This delicate model is dependent on accurately predicting churn of their paid users.

Goal

Based on available data on former customers (users who terminated the subscription in past), KKBOX would like to predict whether an existing user might churn after his/her subscription expired.

Based on the results, the firm's marketing and/or customer experience teams will intervene to gauge better understanding of the situation by opening up communication channels with such likely-to-churn users and by conducting causal analysis on the recently-acquired feedback before it can rectify the situation by taking appropriate tactical and strategic actions.

For starters, the firm shall conduct surveys (email, phone, etc.) with such users to understand their current satisfaction levels with the service and area of improvement (service fees, product features, content assortment, etc.). The firm shall then leverage the survey data to find major areas of user distress/dissatisfaction and segment users accordingly to later target them with "personalized" incentives (discounted yearly subscription vs. costlier monthly subscription, etc.) and/or marketing activities (using mobile app rather than web app, sending regular and gentle reminders to use the service, educate user on unknown features to enhance their experience, etc.) to reverse their departure intentions (if any).

The marketing team can feed the relevant results of causal analysis to product management team, which can further devise new product features and/or enhance existing ones, thereby strategically enhancing customer experience over time.

Input

The data is available at the Kaggle website and is rich in both length (size) and breadth (attributes).

a) Subscription transaction data, depicting transactions of users up until 2/28/2017. Size/shape: 22mn x 9.

b) Daily user logs, describing listening behaviors of a user; data collected until 2/28/2017. Size/shape: 400mn x 9.

c) Users' basic demographic data refreshed on 11/13/2017. (Note that not every user in the dataset is available). Size/shape: 7mn x 6.

The training dataset comprises of churn behavior of 1mn users during the month of March 2017. Similarly, test dataset has churn results of roughly 1mn users for the month of April 2017.

Output

The output of the process would be user id and a flag indicating whether the user intends to churn in the upcoming month when his/her subscription expires.

Data Dictionary

Sales Transaction

i. msno: user id

ii. payment_method_id: payment method

iii. payment_plan_days: length of membership plan in days

iv. plan_list_price: in New Taiwan Dollar (NTD)

v. actual_amount_paid: in New Taiwan Dollar (NTD)

vi. is_auto_renew

vii. transaction_date: format %Y%m%d

viii. membership_expire_date: format %Y%m%d

ix. is_cancel: whether the user canceled the membership in this transaction

Usage

i. msno: user id

ii. date: format %Y%m%d

iii. num_25: # of songs played less than 25% of the song length

iv. num_50: # of songs played between 25% to 50% of the song length

v. num_75: # of songs played between 50% to 75% of the song length

vi. num_985: # of songs played between 75% to 98.5% of the song length

vii. num_100: # of songs played over 98.5% of the song length

viii. num_unq: # of unique songs played

ix. total_secs: total seconds played

Demographic

i. msno: user id

ii. city

iii. bd: age. Note: this column has outlier values ranging from -7000 to 2015, please use your judgement.

iv. gender

v. registered_via: registration method

vi. registration_init_time: format %Y%m%d

Process

1) Analyze data and conduct EDA: User logs and transaction data would need to be aggregated at some level before exploratory data analysis could be done. This would be challenging due to the sheer size of the data.

2) Feature engineering: User logs and transaction data would need to be aggregated at user id level and denormalized, before it could be plugged into the classification model. Since the data size is abnormally large, aggregation would pose some challenges, considering the limited CPU and memory resources availability on my PC.
3) Train model: As the requirement is to predict churn i.e. whether a user is likely to leave the subscription service or not, a binary classification model needs to be deployed. It would be better to start off with Logistic Regression, and then with KNN, Random Forest and Support Vector Machines (SVM).
4) Test model: After comparing the classification results of all the models on test data, the model with least classification error rate will be selected as the optimal one.

Challenges

The size of usage data is in tune of 400mn records, which might pose difficulty in repeatedly aggregating the data for EDA and for building a suitable denormalized feature set for a data science algorithm to work upon. Due to limited memory resources on my PC, it would be impossible to ingest the complete dataset and play around with it.

I plan to overcome this issue as follows:

1) Use turicreate package to avoid memory constraints, as SFrame dataframe object of the package is stored column-wise on persistent storage (e.g. disk).
2) Select a random subset of users from the training dataset and extract corresponding transactions and usage logs to aggregate later, thereby making EDA, feature engineering and model training cycles faster.