December 10th, 2017

**MIS-637-A (Fall 2017)**

**PROJECT PROPOSAL – KKBOX STREAMING SERVICE CUSTOMER CHURN**

**PROBLEM STATEMENT**

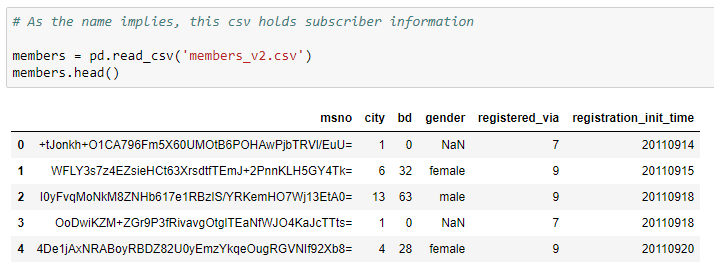
KKBox (herein referred to as “the client”) has been at the forefront of the Asian music streaming market from when it was founded in 2004 up until 2014 when YouTube and Apple entered their market. The client is experiencing increasingly disturbing customer attrition and needs to put a handle on it before irreparable damage is done to their bottom line.

The client has published a sample of anonymized customer information, and one month’s worth of transactions and listening behavior. They want valuable information on customer churn mined from this data to aid their decision-making going forward.

**SAMPLE DATA** - [Link](https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data)

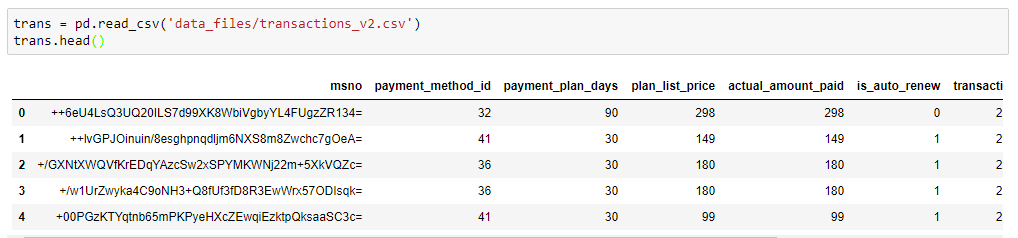
Customer information: 6 features (user-id, city code, age, gender, registration channel, registration time) and 795,091 rows of observations.

members.csv



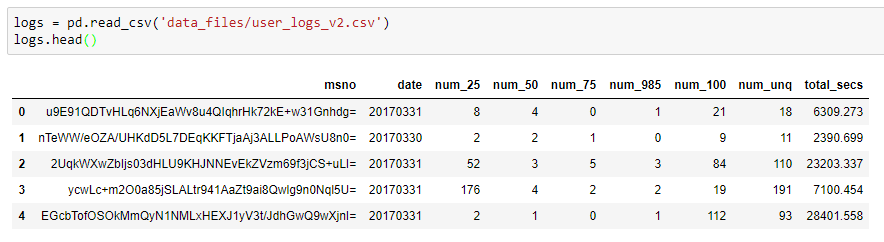
Transactions: 9 features (user-id, payment method, payment plan, plan price, amount paid, auto renew or not, transaction date, membership expiration date, is a cancel or not) and 1,431,010 rows.

transactions.csv



User Logs: Logs of daily listening behavior for one month 9 features (user-id, date, number of songs listened quarter of the way, 50% of the way, 75% of the way, 98.5% of the way, 100% of the way, number of unique songs listened to that day, total seconds of songs consumed that day) and 18,396,363 rows.

user\_logs.csv



**METHODOLOGY**

Following the CRISP-DM, with descriptive discovery run on the individual datasets and combined, null value imputation methods applied accordingly, I plan to run scikit-learn’s CART decision tree algorithm on a sample of the data and predict the probability of churn. I will then measure the log loss of my results against the client’s prediction benchmark.

**SOFTWARE**

I will explore the data with a jupyter notebook running Python 2.7, combine the data with Alteryx, and train the respective algorithms using the scikit-learn data science package.