官网的数据集介绍 方便查询

train.csv

the train set, containing the user ids and whether they have churned.

* msno: user id
* is\_churn: This is the target variable. Churn is defined as whether the user did not continue the subscription within 30 days of expiration. is\_churn = 1 means churn,is\_churn = 0 means renewal.

sample\_submission\_zero.csv

the test set, containing the user ids, in the format that we expect you to submit

* msno: user id
* is\_churn: This is what you will predict. Churn is defined as whether the user did not continue the subscription within 30 days of expiration. is\_churn = 1 means churn,is\_churn = 0 means renewal.

transactions.csv

transactions of users up until 2/28/2017.

* msno: user id
* payment\_method\_id: payment method
* payment\_plan\_days: length of membership plan in days
* plan\_list\_price: in New Taiwan Dollar (NTD)
* actual\_amount\_paid: in New Taiwan Dollar (NTD)
* is\_auto\_renew
* transaction\_date: format %Y%m%d
* membership\_expire\_date: format %Y%m%d
* is\_cancel: whether or not the user canceled the membership in this transaction.

user\_logs.csv

daily user logs describing listening behaviors of a user. Data collected until 2/28/2017.

* msno: user id
* date: format %Y%m%d
* num\_25: # of songs played less than 25% of the song length
* num\_50: # of songs played between 25% to 50% of the song length
* num\_75: # of songs played between 50% to 75% of of the song length
* num\_985: # of songs played between 75% to 98.5% of the song length
* num\_100: # of songs played over 98.5% of the song length
* num\_unq: # of unique songs played
* total\_secs: total seconds played

members.csv

user information. Note that not every user in the dataset is available.

* msno
* city
* bd: age. Note: this column has outlier values ranging from -7000 to 2015, please use your judgement.
* gender
* registered\_via: registration method
* registration\_init\_time: format %Y%m%d
* expiration\_date: format %Y%m%d

Data Extraction Details

One important information in the data extraction process is the definition of membership expiration date. Suppose we have a sequence for a user with the tuple of (transaction date, membership expiration date, and is\_cancel):

(2017-01-01, 2017-02-28, false)

(2017-02-25, 0217-03-15, false)

(2017-04-30, 3017-05-20, false)

(data used for demo only, not included in competition dataset)

This user is included in the dataset since the expiration date falls within our time period. Since the subscription transaction is 30 days away from 2017-03-15, the previous expiration date, we will count this user as a churned user.

Let's consider a more complex example derive the last one, suppose now a user has the following transaction sequence

(2017-01-01, 2017-02-28, false)

(2017-02-25, 2017-04-03, false)

(2017-03-15, 2017-03-16, true)

(2017-04-01, 3017-06-30, false)

The above entries is quite typical for a user who changes his subscription plan. Entry 3 indicates that the membership expiration date is moved from 2017-04-03 back to 2017-03-16 due to the user making an active cancellation on the 15th. On April 1st, the user made a long term (two month subscription), which is 15 days after the "current" expiration date. So this user is not a churn user.

Now let's consider the a sequence that indicate the user does not falls in our scope of prediction

(2017-01-01, 2017-02-28, false)

(2017-02-25, 2017-04-03, false)

(2017-03-15, 2017-03-16, true)

(2017-03-18, 2017-04-02, false)

Note that even the 3rd entry has member ship expiration date falls in 2017-03-16, but the fourth entry extends the membership expiration date to 2017-04-02, not between 2017-03-01 and 2017-03-31, so we will not make a prediction for the user.