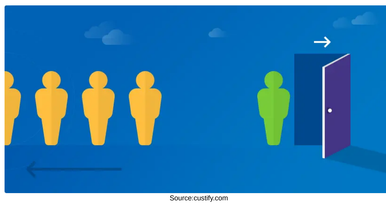
What is the Problem?

WSDM - KKBox's Churn Prediction Challenge

What is Churn ?



Source:custify.com

Churn quantifies the number of customers who have left your brand by cancelling their subscription or stopping paying for your services.

Source:https://www.appier.com/blog/churn-prediction/

Business Impact

Since the music streaming service providers are becoming
more competitive day by day one of the major problem these companies are
facing is customer retention.
High Churn Rate is bad news for any business as it costs five times as much to attract a
new customer as it does to keep an existing one.
A high customer churn rate will hit your company's finances hard.

kkbox.com

Prbolem Statement

To predict whether a user will churn after his/her subscription expires.
Specifically, we want to forecast if a user make a new service subscription transaction within 30 days after the current membership expiration date.

Client

KKBOX is Asia's leading music streaming service,
holding the world's most comprehensive Asia-Pop music library with over 30 million tracks,
supported by advertising and paid subscriptions.

USeful Links

- https://www.kaggle.com/c/kkbox-churn-prediction-challenge
- https://arxiv.org/ftp/arxiv/papers/1802/1802.03396.pdf
- https://www.appier.com/blog/churn-prediction/

Business objective and constraints

The company uses survival analysis techniques to determine the residual membership life time for each subscriber.
By adopting different methods, KKBOX anticipates they'll discover new insights to why users leave
Accurately predicting churn is critical to long-term success.
Even slight variations in churn can drastically affect profits.
No latency connstrains
Minimum error
Have probabilistic Output

Machine Learning Problem

Description

The churn/renewal definition can be tricky due to KKBox's subscription model. Since the majority of KKBox's subscription length is 30 days, a lot of users re-subscribe every month. The key fields to determine churn/renewal are transaction date, membership expiration date, and is_cancel. Note that the is_cancel field indicates whether a user actively cancels a subscription. Subscription cancellation does not imply the user has churned. A user may cancel service subscription due to change of service plans or other reasons. The criteria of "churn" is no new valid service subscription within 30 days after the current membership expires.

The training and the test data are selected from users whose membership expire within a certain month. The train data consists of users whose subscription expires within the month of February 2017, and the test data is with users whose subscription expires within the month of March 2017. This means we are looking at user churn or renewal roughly in the month of March 2017 for train set, and the user churn or renewal roughly in the month of April 2017. Train and test sets are split by transaction date,

source:visualstudiomagazine.com

Data

train_v2.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_v2.csv

- 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.

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
- gender
- registered_via: registration method
- registration_init_time: format %Y%m%d

Sample DataPoints

Train.csv

```
      msno    is_churn
0   waLDQMmcOuZjLDaV1ddDkgCrB/jl6sD66Xzs0Vqax1Y=    1
1    QA7uiXy8vIbUSPOkCf9RwQ3FsT8jVq2OxDr8zqa7bRQ=    1
2    fGwBva6hikQmTJzrbz/2Ezjm5Cth5jZUNvXigKK2AFA=    1
3   mT5V8rEpa+8wuqi6x0DoVd3H5icMKkE9Prt49UlmK+4=    1
4   XaPhtGLk/5UvvOYHcONTwsnH97P4eGECeq+BARGItRw=    1
```

Members.csv

```
                        msno                          city    bd  gender  registered_via
registration_init_time
0   waLDQMmcOuZjLDaV1ddDkgCrB/jl6sD66Xzs0Vqax1Y=        18.0    36.0    female   9.0
20050406.0
1    QA7uiXy8vIbUSPOkCf9RwQ3FsT8jVq2OxDr8zqa7bRQ=        10.0    38.0    male     9.0
20050407.0
2    fGwBva6hikQmTJzrbz/2Ezjm5Cth5jZUNvXigKK2AFA=        11.0    27.0    female   9.0  2
0051016.0
3   mT5V8rEpa+8wuqi6x0DoVd3H5icMKkE9Prt49UlmK+4=        13.0    23.0    female   9.0  2
0051102.0
4   XaPhtGLk/5UvvOYHcONTwsnH97P4eGECeq+BARGItRw=         3.0 27.0         male     9.0
20051228.0
```

Posing as a Machine Learning Problem

Binary class classification : is_churn either 0 or 1

Evaluration Metrics

LogLoss: KPI
Other metric to keep track
Confusion Matrix
F1 Score

USeful Blogs And Reference

- https://www.geeksforgeeks.org/python-working-with-date-and-time-using-pandas/
- https://www.kaggle.com/c/kkbox-churn-prediction-challenge/discussion/46078
- https://www.kaggle.com/jeru666/did-you-think-of-these-features
- https://arxiv.org/ftp/arxiv/papers/1802/1802.03396.pdf

Next

Exploratory Data Analysis

What do you want me to stay - EDA-KKBOX

In [ ]:

In [ ]:

In [ ]:

In [ ]: