Proposal :

Project :

Customer churn modelling for a music company.

https://www.kaggle.com/c/kkbox-churn-prediction-challenge

Domain background :

Customer Churn is a metric used to quantify the number of customers who left the company. For SaaS businesses, it can be defined by those who unsubscribed or canceled the service contracted earlier.

Customer Churn can be crucial to evaluate customer satisfaction over periods of time, especially when measuring negative impacts which recent changes on the website, product supplier and others may have caused customers to leave.

Churn is especially relevant in contractual circumstances, which are often referred to as a "subscription setting," as cancellations are explicitly observed

Since the KK Box model is subscription based model, customer churn is really critical metric for them.

In summary, customer churn or Churn Rate serves as a thermometer showing if customers are having a good experience with the products and services offered. Because otherwise, they will leave.

But how can companies predict Customer Churn? there is no correct answer to this question. It is believed customers behave similarly in a certain way when they are unsatisfied and about to leave. The signals can be plenty like not replying to email marketing, not logging in, not searching for products, by filing a complaint on the website, by living in a region where the competitor got suddenly stronger etc.

That’s why machine learning is becoming highly sought after to detect Customer Churn. By showing the machine positive cases of churn the model calculates the probability that someone is about to leave. Identifying those customers at risk, the company can send an especial offer to engage them back to the store. It could come in form of a discount coupon or a warm-up message to remember the customer about their importance to the company

Relevant Links : https://medium.com/towards-data-science/churn-prediction-with-machine-learning-e6612cd5538f

https://www.datascience.com/blog/what-is-a-churn-analysis-and-why-is-it-valuable-for-business

Problem Statement :

The 11th ACM International Conference on Web Search and Data Mining (WSDM 2018) is challenging us to build an algorithm that predicts whether a subscription user will churn using a donated dataset from KKBOX. WSDM (pronounced "wisdom") is one of the the premier conferences on web inspired research involving search and data mining. They're committed to publishing original, high quality papers and presentations, with an emphasis on practical but principled novel models.

For a subscription business, accurately predicting churn is critical to long-term success. Even slight variations in churn can drastically affect profits.

KKBOX is Asia’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.

In this competition we are tasked to build an algorithm that predicts whether a user will churn after their subscription expires. Currently, 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 so they can be proactive in keeping users dancing.

In this task, we will be predicting whether a user will churn after their subscription expires. Specifically, we want to see if a user make a new service subscription transaction within 30 days after their current membership expiration date.

As a music streaming service provider, KKBox has members subscribe to their service. When the subscription is about to expire, the user can choose to renew, or cancel the service. They also have the option to auto-renew but can still cancel their membership any time.

Problem statement Link : https://www.kaggle.com/c/kkbox-churn-prediction-challenge

Datasets and Inputs :

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. Note that a 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 train 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, as well as the public and private leaderboard data.

In this dataset, KKBox has included more users behaviors than the ones in train and test datasets, in order to enable participants to explore different user behaviors outside of the train and test sets. For example, a user could actively cancel the subscription, but renew within 30 days.

Tables

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 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.

Source : https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data

Solution Statement :

The solution to this problem is to build a churn model algorithm which predicts wether a given user will churn in the next 30 days or not. The metrics that is used for benchmarking the model is the log loss function.

We will be using an xgboost model to calculate the probability of the churn of a given user. We will be making use of the train, user_logs and transactions data and members data to train the model.

Benchmark Model :

For the benchmark model, we will be simply using the mean of the training dataset and apply that to the test dataset. Since in the training set we have the indicator that customer has churned, the mean value is then added to the test dataset and log loss function is used to calculate the score.

Using this approach, the baseline score we get is **0.30786**

Since this is a really naive model, we expect an improvee in accuracy as we refine our model

Evaluation Metrics :

The evaluation metric for this competition is

Log Loss

$$logloss=-1N\sum i=1N(yi\log(pi)+(1-yi)\log(1-pi))$$

where N is the number of observations, *log* is the natural logarithm, *yi* is the binary target, and *pi* is the predicted probability that *yi* equals 1.
Note: the actual submitted predicted probabilities are replaced with $max(min(p,1-10-15),10-15)$.

Submission File

For each user id (msno) in the test set, you must predict the probability of churn (a number between 0 and 1). The file should contain a header and have the following format:

msno,is_churn

ugx0CjOMzazClkFzU2xasmDZaoIqOUAZPsH1q0teWCg=,0.5

zLo9f73nGGT1p21ltZC3ChiRnAVvgibMyazbCxvWPcg=,0.4

f/NmvEzHfhINFEYZTR05prUdr+E+3+oewvweYz9cCQE=,0.9

Link : https://www.kaggle.com/c/kkbox-churn-prediction-challenge#evaluation

Project Design :

The initail state will improve data exploration and cleanup. We have to look for outliers and missing data.

For e.g. birth age has a value of -7000 to 2015 and needs to be treated.We also need to treat the outliers of the respective data.

Another part is that there are 60% of the gender values are NULL.

These are observations from basic exploration. We need to go further into the data exploration to get deep insights from the data.

Now since we need to include only those user data for which we have inforamtion in the train data set. So we need to combine the transaction and members dataset with train dataset. After combining we need to treat the data accordingly and check for outliers. We might also need to work around duplicate records and we might drop the previous transactions of an user(not sure of that till now)(we might also go for a time series anlaysis for further evaluation).

Now comes the user logs.Since the user log is a gigantic 40 gb data, we need to work out the data properly. We might need to trim the data to include only latest records rather than going from start of time.

Now we need to create a prediction model.Since the data is large, we might go for xgboost or lgbm model to cut down the running times. Differnet iteration of the model has to be tested to get the best score.The evaluation metric to be used is logloss.

Now our model is trained, the next thing we need is to test the model and make further adjustment to get the desired accuracies.

Our new model should be an improvemnet on our benchmark model score.