Use this format (Don't remove Project 0)

Project # 0

Group members:

Chang, Dong (20429617)

Xiao, Wang (20532766)

Yuhao, Xie (20593455)

Title: Kaggle Competition : WSDM - KKBox's Churn Prediction Challenge

Description:

Introduction : KKBOX, one of the leading pop music platform in Asia, has set up this competition which requires an algorithm to predict if users from a given dataset (or any with the same format) will continue their current subscription on KKBOX. As also KKBOX is known as a leading stream platform in Asia, which is supported by both advertising and user subscriptions.

Timeline : Based on the Timeline from the competition :

September, 18th, 2017: Competition begins

December 10, 2017: Team Merger Deadline

December, 17th, 2017: Competition Ends

December, 19th, 2017: Winner announcement

January, 9th, 2018: Workshop paper submission deadline

February, 8th, 2018: WSDM cup workshop

All deadlines are at 11:59 PM UTC on the corresponding day unless otherwise noted. The competition organizers reserve the right to update the contest timeline if they deem it necessary.

For our team , we have a basic timeline following as below :

Oct, 1 , 2017: Team formation and start of the project (weekly meeting basis)

Oct, 5 , 2017: First Team meeting with role and language setup

Oct, 13 , 2017: Second Team meeting and start working on the general algorithm(functions, formulas)

Oct, 20 , 2017: Third Team meeting : optimize the algorithms with a brief testing case (not coded,yet) with different sample sizes

Oct, 27 , 2017: Fourth Team meeting : upon the results and conclusions from the previous week, start coding or keep working on algorithms

Nov, 22 , 2017: Deadline on coding section and testing step 1

Nov, 30 , 2017: Submit the final results with documentations(if any)

December, 2 , 2017: Competition Ends with Final presentation (TBD)

Sample Submission:

msno,is_churn

ugx0CjOMzazClkFzU2xasmDZaoIqOUAZPsH1q0teWCg=,0.5

zLo9f73nGGT1p21ltZC3ChiRnAVvgibMyazbCxvWPcg=,0.4

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

Etc.

Note : In this sample submission, the first column is the name of the existing user, which can be found in the given .csv file. The second column states the predicted probability of the given user will churn(withdraw the sub).

Evaluation methodology :

The evaluation metric for this competition is **Log Loss**

$$logloss = -\frac{1}{N}\sum_{i-1}^{N}\left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right)$$

where N is the number of observations, log is the natural logarithm, y_i is the binary target, and p_i is the predicted probability that y_i equals 1.

where N is the number of observations, *log*log is the natural logarithm, $y_i$yi is the binary target, and $p_i$pi is the predicted probability that $y_i$yi equals 1.

Note: the actual submitted predicted probabilities are replaced with *max*(*min*(*p*,1−10−15),10−15)max(min(p,1−10−15),10−15) .

First attempt : Observe the raw data set :

Observe that in .csv, our data are separated by commas and the second column stands a probability between 0 and 1

1	ids,label
2	0,8
3	1,6
4	2,1
5	3,1
6	4,6
7	5,5
8	6,2
9	7,2
10	8,9
11	9,7
12	10,9

Quick Reference :

Check the link for the competition :

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