KKBox's User Stickiness Analysis

Yujia Wang, Liyi Cao, Tong Ye, Weixuan Jiang

yujia8@bu.edu, caoliyi@bu.edu , tongyewh@bu.edu, jwx0728@bu.edu

1. Project Task

KKBOX is an Asia's leading music streaming service and it holds tens of millions of Asia-Pop musics and since it provides services to millions of people, they need a model on accurately predicting churn of their paid users.

So our input is the data provided by Kaggle, and the task is to build an algorithm that predicts whether users will be lost after the subscription expires. And to analyze the reason for the user leaving in order to be proactive in keeping users by using different methods.

2. Related Work

Currently, the company uses survival analysis techniques to determine the residual membership life time for each subscriber. Survival analysis is a statistical method that considers the results and considers the survival time. It can make full use of the incomplete information provided by the censored data to describe the distribution characteristics of survival time and analyze the main factors affecting the survival time.

3. Approach

Logistic Regression:Logistic Regression is a basic classification method taught in class. Basically, it uses a function to model a binary variable.

$$g(z) = \frac{1}{1+e^{-z}} \qquad h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

And when $h_\theta(x)$> 0.5, we predict it as 1;else we predict it as 0.

SVM:A Support Vector Machine (SVM) is a discriminative classifier defined by a separating hyperplane. The input of SVM is training data and their labels(supervised learning), the output is an optimal hyperplane which could categorizes test data for classification.

Neural Network:Neural networks are a set of artificial neurons like the human brain, which are designed to recognize patterns. They interpret sensory data through Neural network, labeling or clustering raw input.

4. Dataset and Metric

train.csv contains the IDs and whether these users have churned or not.

transactions.csv gives us details like *payment method* or whether the subscription was cancelled.

user_logs.csv contains the listening behaviour of a user in terms of number of songs played.

members.csv includes the user's age, city, and such for users that have these membership information.

sample_submission_zero.csv serves as the *test* data set for the users for which we are tasked to predict their behaviour.

5. Preliminary Results

So far, we have analyzed the content of the dataset. Initially identified the types of algorithms we will use in the project and developed our milestones and timelines. We also found articles and books related to our project, as well as the abbreviations of existing algorithms.

6. Timeline and Roles

Note, each teammate should be assigned some non-trivial coding task.

Task	Deadline	Lead
Data collection and cleaning	11/04/18	Liyi Cao
Implement logistic regression	11/14/18	Weixuan Jiang
Implement Support Vector Machine	11/24/18	Tong Ye
Implement Neural Network	12/04/18	Yujia Wang
Prepare report and presentation	12/11/18	all

References

1) Peng, Chao-Ying Joanne. "*Logistic Regression.*"
2) Joachims, Thorsten. *Making large-scale SVM learning practical*. No. 1998, 28. Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.
3) Hagan, Martin T., et al. *Neural network design*. Vol. 20. Boston: Pws Pub., 1996.