Machine Learning Engineer Nanodegree

Capstone Proposal

Eakalak Suthampan
November, 2017

WSDM - KKBox's Churn Prediction Challenge

Can you predict when subscribers will churn?

Domain Background

Customer churn prediction is essential for businesses. Early prediction of customers churn can help businesses to propose marketing campaign to prevent their customers from canceling their subscription, product or service. There are two broad approaches for churn analysis. Machine learning which we will focus and survival analysis. Machine learning methods, specifically classification, are widely used due to their high performance and ability to handle complex relationships in data. On the other hand, survival analyses can provide value by answering a different set of questions. Quantities, such as survival and hazard functions, can be used to forecast which customers will churn in a particular time period.

This capstone will take the problem from the Kaggle competition "WSDM - KKBox's Churn Prediction Challenge". The KKBOX is Asia's leading music streaming service and would like 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.

Problem Statement

The objective for the Kaggle competition "WSDM - KKBox's Churn Prediction Challenge". is predicting whether a user will churn after their subscription expires. **The criteria of "churn" is no new valid service subscription within 30 days after the current membership expires.**

Datasets and Inputs

There are datasets as follows.

1. train.csv (44.5 MB) is the train set, containing the user ids and whether they have churned. The train data consists of users whose subscription expires within the month of February 2017.
2. sample_submission_zero.csv (43.5 MB) is the test set, containing only user ids. The test data is with users whose subscription expires within the month of March 2017.
3. transactions.csv (1.61 GB) is transactions of users up until 2/28/2017.
4. user_logs.csv (28.4 GB) describing listening behaviors of a user. Data collected until 2/28/2017.
5. members.csv (352 MB) consist of user information. Note that not every user in the dataset is available.

As the day I wrote this proposal, there were leaks of the labels for the test set sample_submission_zero.csv. Kaggle said they will have new updated datasets soon. Please see detail at the Kaggle competition "WSDM - KKBox's Churn Prediction Challenge".

Solution Statement

- Build binary classification model to predict whether a user will 'churn' or 'not churn' after their subscription expires.
- The training label of the classification is the column 'is_churn' on the train.csv.
- The classification will based on the various features on members.csv, transactions.csv, user_logs.csv.
- I will randomly split 10% of the train.csv to be use as testing data since I don't know the actual labels of sample_submission_zero.csv which is used as test set in the competition.

Benchmark Model

For the benchmark, I will use simple prediction by

- From user_logs.csv, if user has no activity longer than the last 30 days, then predict this user as churn.
- May be I will also use logistic regression as benchmark to be compared with the ensemble method.

Evaluation Metrics

Since we would like to identify customer churn as much as possible (high recall) while still getting acceptable precision. Therefore, I will use f1_score as evaluation metric since the f1_score conveys the balance between the recall and the precision .

By the way, the competition use logloss as evaluation metric. logloss is more precise to measure classification performance because it is based on prediction in probabilities. I think logloss is more appropriate when using in the competition but I would like to use f1_score because I think it is more intuitive to explain to business in term of recall, precision.

Project Design

- First, the user_log.csv is too large (29 GB) to be processed in a normal way so may be I need to split it to many chunks then process and aggregate them.
- Feature engineering on the user_log.csv and transactions.csv to create new useful features such as
  - duration of membership for each user
  - number of subscription renewals in the past for each user
  - number of activities for each user
  - number of activities in last 30 days for each user
  - total seconds played for each user
- Since it's normal that the number of 'churn' will much less than the number of 'not churn'. I plan to use ensemble model such as Random Forest or XGBoost to handle the imbalance classification [4].
- For handle imbalance classification, I'm still not sure whether I should do oversampling, undersampling or SMOTE [4] to balance the label classes or not.
- Initial models evaluation between Random Forest, XGBoost to find the best algorithm.
- use gridsearchcv to tune parameters of the chosen algorithm.
- explain the results of the evaluation metrics (f1_score, recall, precision, logloss)
- explain feature importance.

References

[1] "WSDM - KKBox's Churn Prediction Challenge" https://www.kaggle.com/c/kkbox-churn-prediction-challenge

[2] S. Figini, Customer relationship: a survival analysis approach, (to appear in Proceedings of Compstat, Roma), 2006. https://www.researchgate.net/profile/Silvia_Figini/publication/228435285_Customer_relationship_a_survival_analysis_approach/links/00463521deec89fae8000000/Customer-relationship-a-survival-analysis-appr pdf

[3] XGBoost https://github.com/dmlc/xgboost

[4] How to handle Imbalanced Classification Problems in machine learning? https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/