Unit 17.2.3 Relax inc. Challenge: Report

1. Data

One dataset was log data of signing in the product. It had 207,917 log data about 8,823 users and did not have any missing values or duplicate values. Another dataset included 12,000 users' information. It had 10 columns (2 date-related, 8 categorical). That is, 3,177 users did not sign in in this period. Two of the columns had missing values, but they were actually empty on purpose (the empty meant that there was no appropriate value).

2. Data Wrangling and EDA

An "adopted user" was defined as a user who had logged into the product on three separate days in at least one seven-day period, and the status of each user was assigned. As the result, 13.8% of the users was the "adopted users".

Each column of the user dataset was prepared and carefully investigated to know if the column had correlations with the status. As the result, 1) newer users tended to be inactive, 2) users who recently logged in were active, 3) email domains and their creation source had different tendencies depending on the account status.

3. Creating a predictive model

Although many kinds of classifiers could be used for this problem, I tried Random Forest and Gradient Boosting Decision Tree (GBDT) because many features were categorical and the classifiers often provided high performance to them.

The preprocessed data were separated into a train set and a test set. The train set was used to find the best combination of the hyper-parameters and train the best model. Then, the evaluation of the model was conducted on the test set. To evaluate a model, a metric needed to be chosen. In this case, I chose f1 score as the metric because the data was imbalanced (14% positive). F1 score is more balanced than recall or precision. The best model was GBDT, and the f1 score on the train set was 0.89 and 0.90 on the test set. The two scores were very close and it was said the model was not overfitting. If Relax inc. wants to pick up all potential users who could adopt and doesn't mind that some false positives would be included, it would be one choice to use recall as the metric of the performances because recall represents how many positive samples were labeled as positive.

4. Important factors to predict future user adoption

According to the feature importance of the best model, the most important factors to predict future user adoption were the account age and recent login. Their scores were predominantly higher than the others. Users who recently logged in and users who are long-term customers tend to adopt the product. E-mail domains and whether a user was invited by anyone would also have some impacts on the prediction. For example, users using Gmail or yahoo mail were more likely to be active.