Summary Report for the Relax Challenge

Data Wrangling:
'adopted user' is first computed from the 'user_engagement' dataset, by first sorting by 'user_id' and 'time_stamp', then looping over to check 'user_id' and 'time_stamp' for the adjacent three entries. In addition, missing data are present in two of the columns of the 'user' dataset, and are dealt with when appropriate.

Findings from EDA:
1. the majority of the users are not adopted users (86.2%)
2. most of the users (> 75%) have visited the product no more than once
3. there is a strong correlation between user adoption and their activity timestamp (last activity timestamp and account creation timestamp)
4. there is a weak correlation between 'adpoted_user' and their organization ('org_id'), as well as user creation source

Feature Engineering: two new features are created:

1. 'time_delta': time delta between a user's last session and account creation time
2. 'total_visits': the total number of visits the user has to the product

Results shown that both features strongly correlate to the user adoption of the product

Machine learning modeling result:
Three different types of classification algorithms are experimented:

1. linear Logistic Regression classifier
2. non-linear Support Vector Machine classifier
3. ensembles of tree-based Random Forest classifier

RandomForestClassifier is found to be the best among the three with a f1 score of 0.952 and an auc score of 0.970 on the testset, it also works the best for handling the imbalanced classes. (Due to the imbalanced nature of this problem, f1_score and area_under_ROC_curve are selected instead of the accuracy score for evaluating the classifiers)

Insights:
Both EDA and feature engineering findings, as well as the feature ranking from the best performing RandomForestClassifier suggest that a user's total number of visits to the product and the time difference between a user's account creation and his/her last activity are the two critical factors indicating the user's adoption of the product.
In addition, the organization the user belongs to ('org_id') and the users' account creation source ('ORG_INVITE' most effective among all) also contributes a little.

Recommended actions:
1. Ways to attract user's visit to the product: emails, news, outreaches, meetups, ads about deals, updates, demos on the product, etc.
2. Ways to attract targeted organizations to invite more users: discount, enterprise pricing, incentives, etc.

Jing Zhao – Nov 2018