Relax Challenge Summary Findings

The user table had 10 features for 12000 users and had some missing values. Using the engagement table, I found 1602 users among all 12000 users (13.35%) are adopted (i.e., logged into the product on 3 separate days in at least one 7 day period) .

EDA

The brief EDA using the user table showed the account creation method (creatio_source) and unix timestamps of last login (last_session_creation_time) are likely to be good predictors in user adoption predictions (see figure 1 and 2). I also plotted other features with adopted user rate, but I could not find any important patterns.

Machine Learning

The models were briefly built and tuned with LightGBM algorithm, a fast gradient boosting method. The performance metric I chose was AP (average precision), a metric that summarizes a precision-recall curve similar to area under precision-recall curve. I chose this metric since I did not want to use a metric that depends on one threshold (e.g., precision, recall, and F1). AUC (area under ROC curve) was another option, but AUC is not appropriate for data with rare positive cases (here 13.35%). The AP score on the test set with the baseline model (without hyperparameters tuning) was 0.7385 and it was increased to 0.7503 after hyperparameter tuning

It was found all features I used had some importances in the baseline model. If hyperparameters are tuned, the model is optimized with lower subsample proportion (subsample=.5), lower number of features (colsample_bytree=.8), and smaller max depth (max_depth=2) for each tree (these reduce overfitting!!). As a consequence, the less important features are ignored and had zero importances (see figure 3).

Although unix timestamps of last login (last_session_creation_time) looks had the highest feature importance, the feature is actually more like a consequence of user adoptions, not a factor of user adoption. In other words, adopted users are more likely to have logged in recently just because they are adopted. Therefore, more useful factors I found here are

- organization id of a user *
- invited user id *
- whether an account is created by invitation to join another user's personal workspace or not
- whether a user is on the marketing email drip or not.

If last_session_creation_time is removed from the model, the model performance is actually very poor. Thus, I would like to recommend to collect more and better information about users to predict user adoptions.

* Note that I kept the categorical columns 'org_id' and 'invited_by_user_id' as numeric predictors since they have too many possible values (417 and 2564, respectively), which are too many for one-hot encoding.

Figure 1

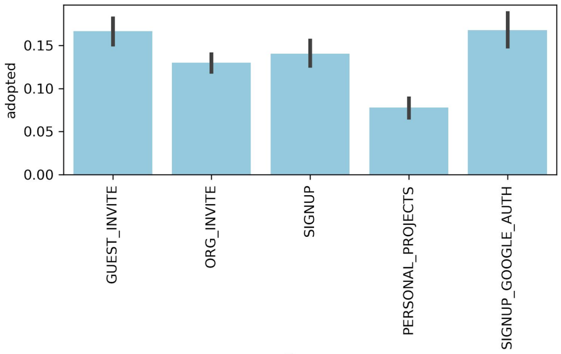


Figure 2

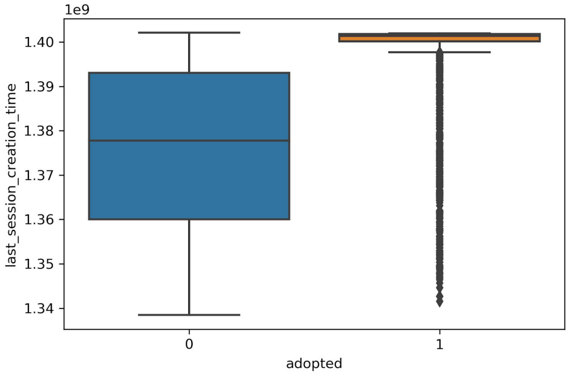
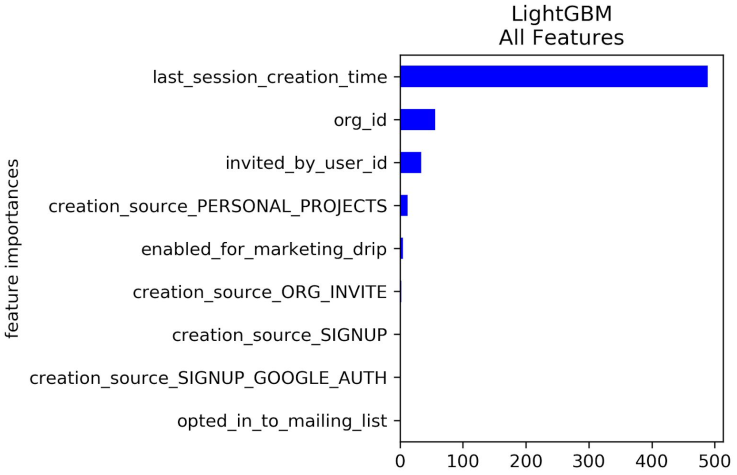


Figure 3



The original question

Defining an "adopted user" as a user who has logged into the product on three separate days in at least one seven day period, identify which factors predict future user adoption. We suggest spending 1-2 hours on this, but you're welcome to spend more or less. Please send us a brief write up of your findings (the more concise, the better no more than one page), along with any summary tables, graphs, code, or queries that can help us understand your approach. Please note any factors you considered or investigation you did, even if they did not pan out. Feel free to identify any further research or data you think would be valuable.