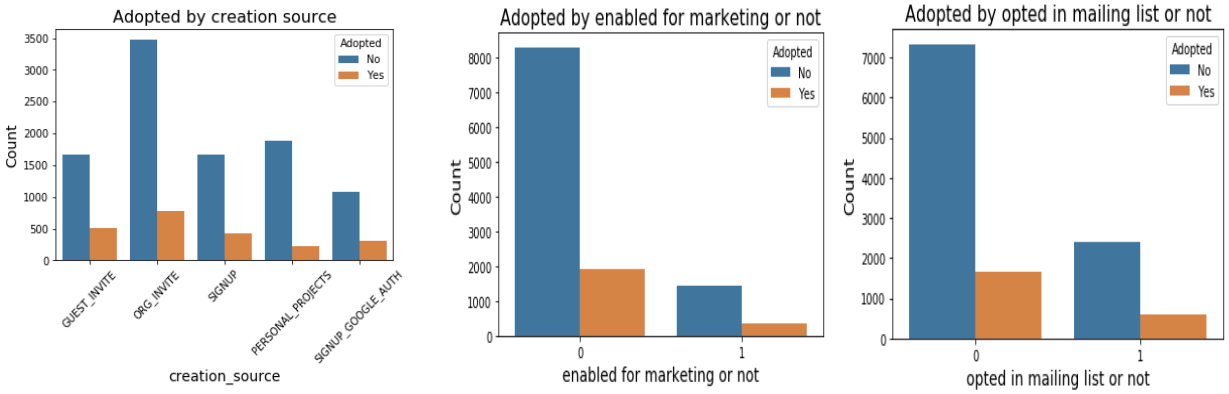
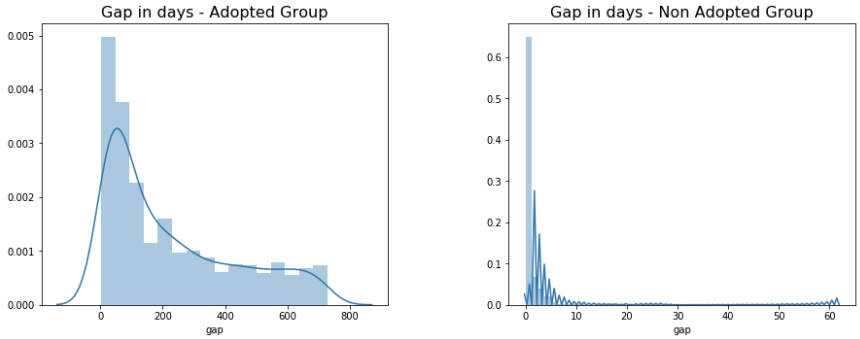
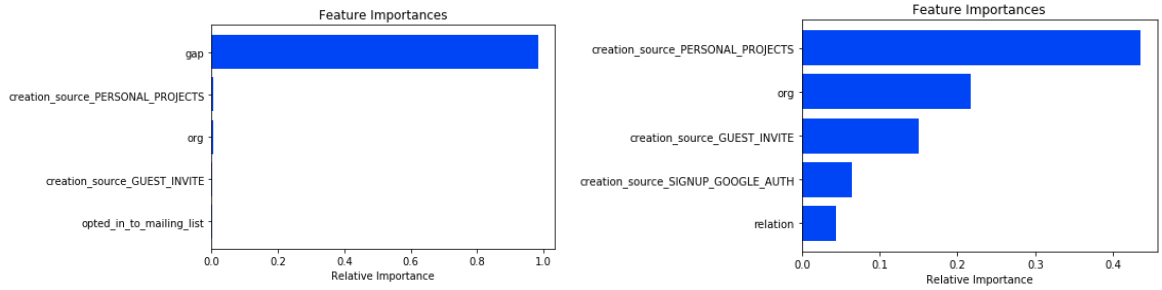
Identify factors that are predictive of future user adoption

Problem Statement:
Given user id, account creation source, user log in information etc. variables, determine which variable is the most predictive of future user adoption.
My Solution:
Step 1: Create a 0-1 target variable “adopted_user” by aggregating login counts from user_engagement.csv file at each user id level. Merge the target variable into the users.csv dataset.
Step 2: Missing value imputation and new feature creations.
Step 3: Data visualization of some explanatory variables within adopted / non-adopted groups.





Step 4: Data processing such as one-hot encoding of categorical features, train test data split to prepare for model building. In this exercise I dropped user names and email etc. text fields as they are not predictive of adoption status by common sense. ID fields are converted to 0-1 features to make it easier to feed into the model. Time columns are converted to “gap” (in days) between last log on time and account creation time.
Step 5: Two Random Forest models are built, one with and one without the field “gap”. Feature importance graphs are generated after fitting each model (left figure: model 1; right figure: model 2), and classification reports such as confusion matrix and AUC scores on the test dataset are also created for each model.



Discussion:

Intuitively, variable "gap" should be closely related to the target variable "adopted_user" because if "gap" is less than 7 days then automatically the adoption status should be 0. This can also be seen from the graph that shows the distribution of "gap" within each adoption status. It also turns out that in the first Random Forest model where we included "gap", its feature importance dominates all other variables and contributes to almost 100%. Model AUC of the first model on the test dataset is 99.8%. In reality this may be considered as data leakage, so the variable "gap" probably shouldn't be used in the model construction, and that's the reason why we are fitting the second model.

In the second model where we don't have feature "gap", its performance is quite poor on both the training and testing data, merely a little better than random guess. Based on the feature importance ranking, it looks like the source of account creation is most predictive in user adoption. Other variables such as organizations the user ID is from, and whether the user is invited by another user, also appeared in the top feature list, but they are not as important as creation source.

More feature engineering can be explored to add extra columns into the dataset, and more models can be tried to compare prediction performances, however, due to the interest of time, these options are not carried out here.

Reference:
https://github.com/lisalb168/relax_challenge