Relax Challenge

In this challenge, I tried to predict retention of users referred to as 'adopted users'. Users are considered adopted users if they logged in 3 times in 7 consecutive days. For the features of the predictive model, I considered the domain of the email address as a series of dummy variables but I decided against this when I realized that there were hundreds of unique domains. I did not want to use too many columns out of fear over-fitting the model. I discarded the invited by user id column and organization id for the same reason but I did create a boolean feature of whether an individual was invited or not. The creation source was turned into 4 dummy columns. Name, creating time, and last login time stamp were dropped.

Visualizing the data (see Figure 1) as bar graphs revealed no noticeable correlations between adopted users and the variables. There were significantly more users were didn't adopt than did. I used gradient boosting because it often gives reliable results for both categorization and regressions. The random hyperparameter search with cross-validation for the gradient boosting model predicted that all users were non adopted users which is not the case. Clearly none of the parameters used had predictive value.I added the total number of logins and number of days between creating time and last session were added to the model. This increased the accuracy to 99%. These two features can only be determined after the fact so they can not predict the behavior of future users.

Its possible that exploring which user invited the user in question may have predictive value. Does being invited by an adopted user make it more likely that a user will also be an adopted user? Including the company id in the model may also have predictive value. What companies are more likely to have adopted users? While this feature is not in the dataset, the reason for creating the account may be useful. If a user creates an account for professional reason than personal ones, is it more likely that they will be an adopted user?

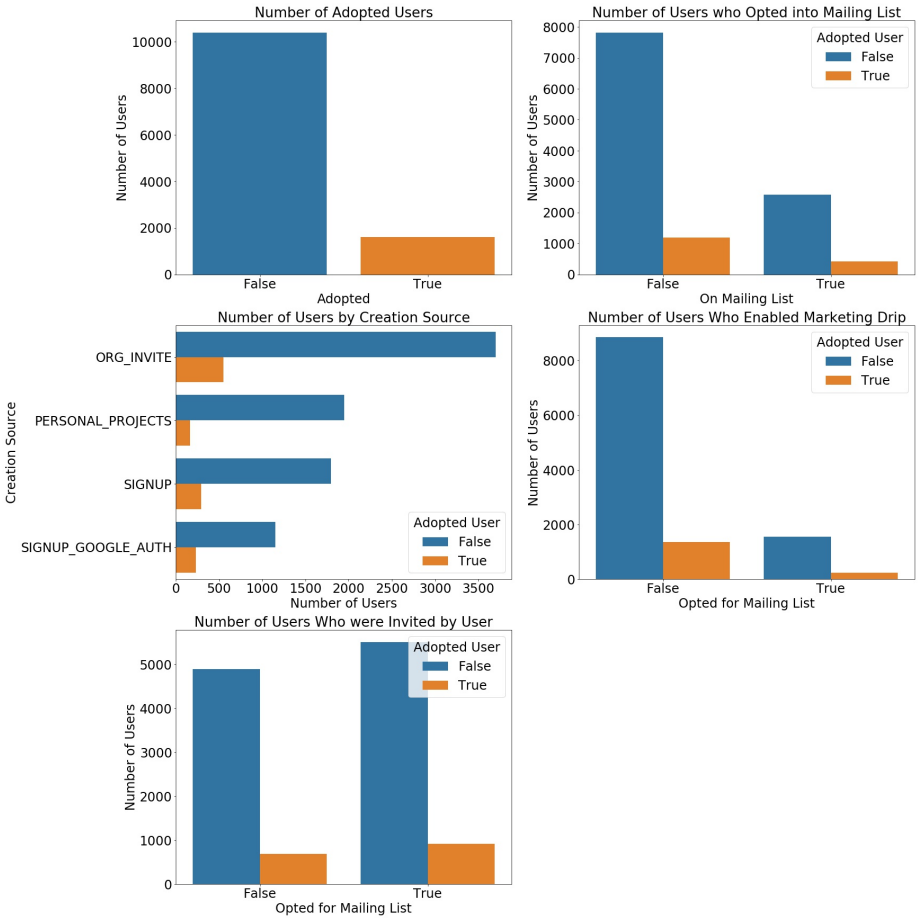

Figure 1: Bar graphs of the features used for predictive models. This does not include the total number of logins or number of days between creating time and last session for the second model.