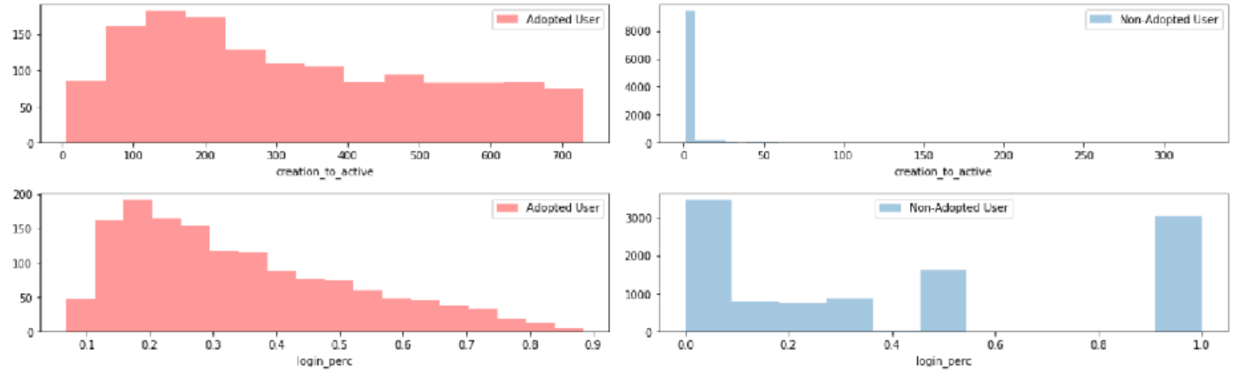
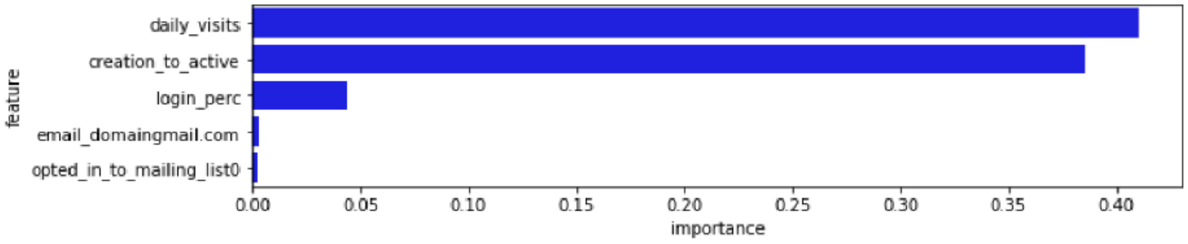
From the 12,000 new users only 1500, or 12% could be considered as adopted-users. More than half of the 12,000 made an account but have no records of ever logging in. There appears to be fake email filters as many of the email domains are fake. There is no history of a user logging in multiple times per day. The dataset is lacking a lot of information about the users as there is only one numerical column.

I first defined 'adopted-user' by resampling and grouping the supplemental data set to determine the users that have history of 3 login days over 7 days. During that resampling period I extracted an additional column 'daily_visits' to load when merging.



Instead of dealing with the outliers manually, I decided to use StandardScaler from sklearn.preprocessing. I did so because the outliers were so frequent in occurrence that I felt it was wrong to hve different specturm of outliers replaced with the same valu.e For categorical variables, I constructed dummy variables for each of the categorical variables. With the preprocssed dataframe, I calculated feature importance using the RandomForestClassifier. Results suggest that the 3 numerical columns have the most impact.


feature
daily_visits
creation_to_active
login_perc
email_domaingmail.com
opted_in_to_mailing_list0
importance