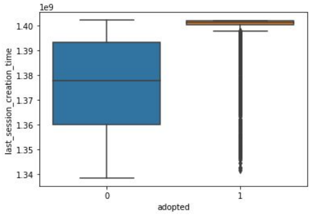
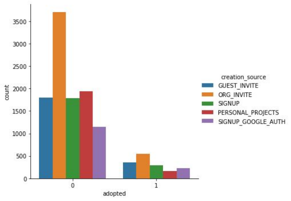
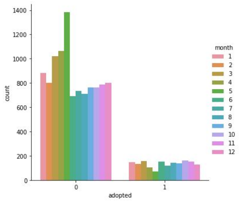
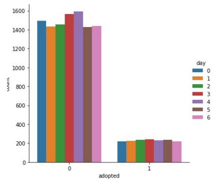
Firstly, I use the usage summary table to find which ones are adopted among total 12000 users. Grouping on ‘user_id’ and rolling the window with a frequency of 7 days gives the counts of logins in 7 days for each user. I get 1607 users out of 12000 who can be considered as ‘adopted’. Add the ‘adopted’ variable in the user table based on above result.

Secondly, explore the data in the user table. Since now we know which users are adopted or which are not, seaborn can be applied to visualize how data is distributed for different groups of users. The boxplot below clearly shows that the ‘adoptive’ users’ last logins are more recent. The bar plot indicates that the creation sources distribute similarly in the 'adopted' users group and the 'unadopted' users group. Both majority users signed up because they are invited to an organization. The second two sources for two groups are both invited to an organization as a guest and signing up via the website.



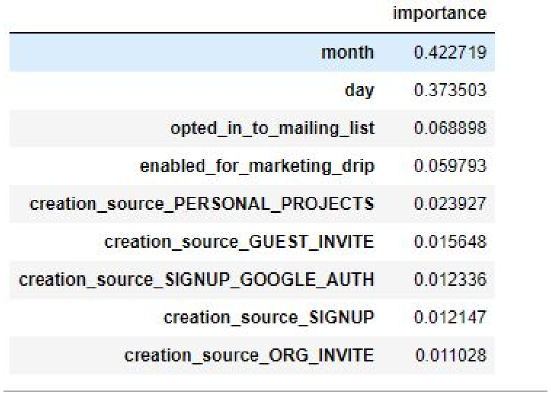






Thirdly, I do some feature selection and engineerings. Many variables in the dataset are not useful to identify whether users are adopted just by their descriptions, such as name,email, and so on. Those will be dropped. Some potential predictive factors are generated from transforming the existed variables, for example month and day (the day of a week) are created from creation_time. I also want to create a factor, the time interval between the creation_time and lass_session_creation_time, but I need some specific information on the creation_time variable, such as what time zone it is referring to, to make sure get the correct time intervals. Hence, I will drop the lass_session_creation_timevariable for now. The third and fourth graphs above are how samples are distributed in different months and weekdays with comparison of adoptive users and not-adoptive users. The ‘month’ distributions are clearly different between the two groups of users, which might be an indicator that month could be a good predictive factor. Then, I encode the selected categorical variables to numeric variables.

Finally, I apply the Random Forest Classifier and get the feature importances as shown below:



	importance
month	0.422719
day	0.373503
opted_in_to_mailing_list	0.068898
enabled_for_marketing_drip	0.059793
creation_source_PERSONAL_PROJECTS	0.023927
creation_source_GUEST_INVITE	0.015648
creation_source_SIGNUP_GOOGLE_AUTH	0.012336
creation_source_SIGNUP	0.012147
creation_source_ORG_INVITE	0.011028

It looks like the factors, `day` and `month` has the most significance on predicting whether users are adopted or not. The five factors related to `creation_source` have the least importance on prediction.