Business problem

We want to identify the “adopted users” from the data and then through analysis, determine the most crucial variables that determine what make a user an “adopted user”.

Data Manipulation and preprocessing

The user engagement file was first read and then, performed a 7 day rolling count grouped by each individual user [1]. This way, adopted users were identified. Then the user data file was read and merged with the adopted user data table. [2]

resampled_data = data.groupby('user_id')['visited'].rolling('7d').count()   [1]

final_df = data1.merge(all_users, left_on = 'object_id', right_on = 'user_id', how = 'left')   [2]

Insights from Data

Roughly 13% of the users are adopted users.

First we observe the creation source variable. We see highest creation activity through guest invites and google auth signups (~**0.17**). Creation through personal projects is significantly low (~**0.08**). This could be because once the users have completed the project, they no longer find any use of the website. The only way they could revisit the site is they think it can help with their personal projects.

The next observed variable is the ‘opted_in_mailing_list’. Only 25% of users have opted to be part of the mailing list. We see no significant differences in percentage between those who signed up and those who didn’t. The percentage of adopted users for these two categories were 13% and 14%.

The next variable to be considered is ‘enabled_for_marketing_drip’. Only 15% of users have signed up for marketing drip. However, 45% of the users who have enabled marketing drip are adopted users as opposed to a paltry 5% for those who did not. Hence, more number of users enabled for marketing drip could potentially scale to higher overall percentage of adopted users. It could be possible that are other users who are not aware of this feature or are aware but don’t exactly know how it works. Perhaps, the website could send an advertisement email informing users about this feature. Then other users may start visiting the website more often.

Next observed variable was the organizations that they belonged to. A dataframe was created to identify the organizations with the highest fractions of adopted users. The top 25 organizations had a fraction of 0.29 or greater. Some of topmost organizations had a greater than 38% fraction of adopted users. This could be one of the telling factors. We must now try and identify what is “common” between these organizations. Are they mostly located in urban/semi-urban areas? Are all of these tech companies? Perhaps, the users from these organizations find matter and products from the website relevant to their work/interests. Do these organizations have more than 5000 employees each? Or, it’s also possible that a combination of these factors drives adopted use rate.

If we had information about the average session time of users, it could help us distinguish between adopted users and those who are not with more brevity. It’s possible that those who aren’t adopted users spent significantly lower durations on the website than those who are adopted users.

Another factor that could have been useful is the users’ level of education. It’s possible that a significant fraction of users are either post graduates or Ph.D graduates. If so, we must try and identify what drives them to the website. We could also identify users whose last created session date was 150 or more days from the registration date. We can then try to identify common traits among these users.