Defining an "adopted user" as a user who has logged into the product on three separate days in at least one sevenday period, identify which factors predict future user adoption. We suggest spending 1-2 hours on this, but you're welcome to spend more or less. Please send us a brief writeup of your findings (the more concise, the better no more than one page), along with any summary tables, graphs, code, or queries that can help us understand your approach. Please note any factors you considered or investigation you did, even if they did not pan out. Feel free to identify any further research or data you think would be valuable.

The ‘adopted user’ is defined as follows

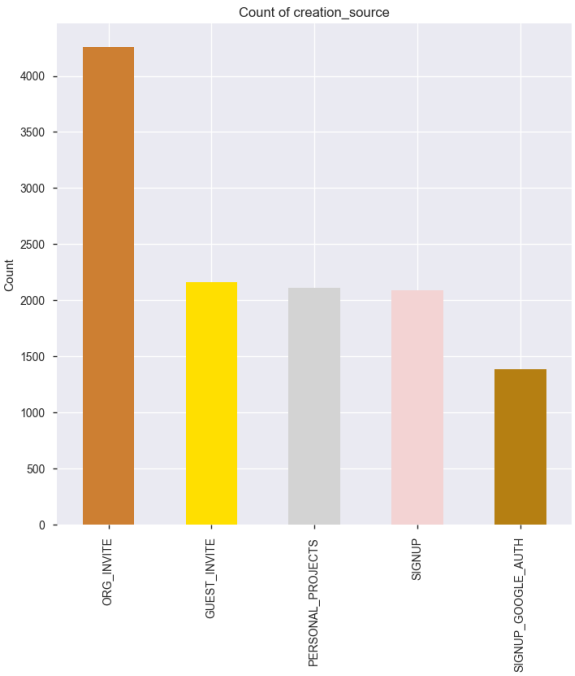
1.Floor the ‘time_stamp’ as day and drop the duplicates; which remove the duplicates if users log on more than once on the same day

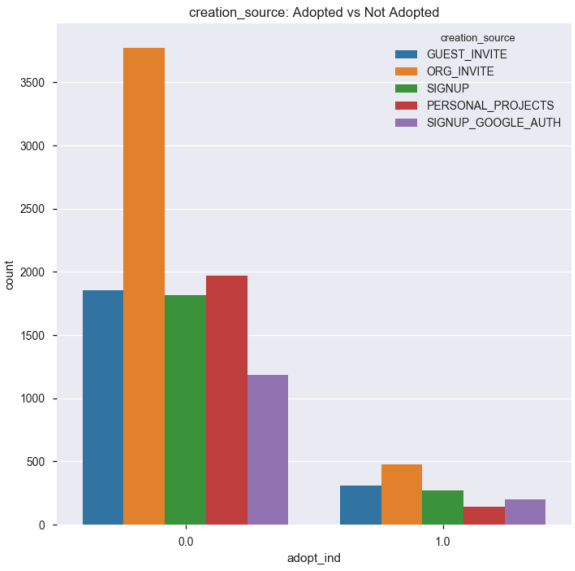
2.Extract the week of ‘time_stamp’ and group by ‘users_id’ and ‘week’

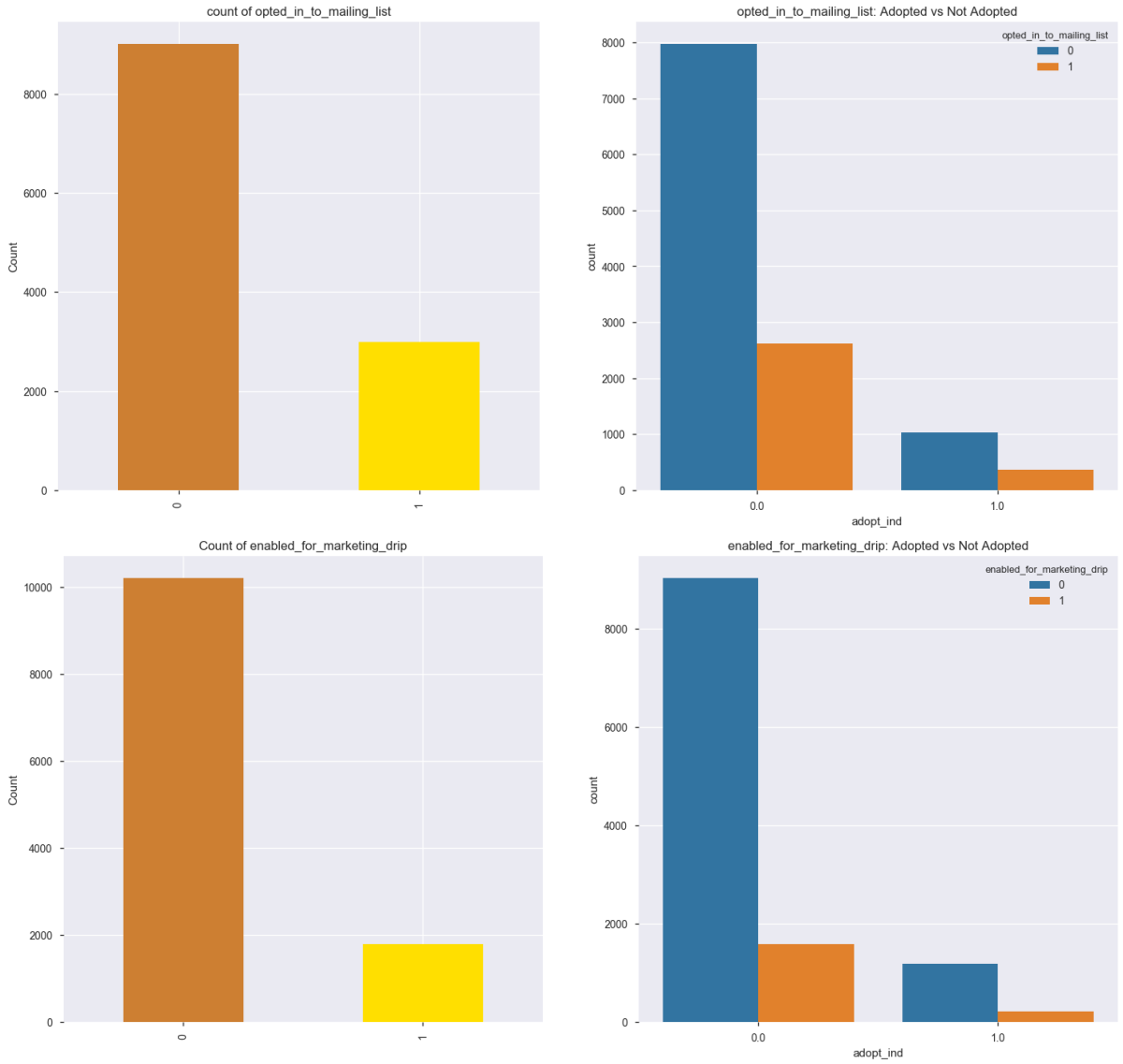
3. Count the frequency and select the one whose frequency bigger than 3

4. Conducting 2 and 3 separately in year 2013 and 2014; then union the selected ‘user_id’

We plot count plots for adopted_users and non- adopted_users according to ‘creation_source’, ‘opted_in_to_mailing_list’, ‘enabled_for_marketing_drip’.







We find that the adopted users and non-adopted users have similar distributions in whether or not enabled for marketing drip and opted in to mailing list'. A slight distributions difference is observed in creation sources: compared with non-adopted users, the adopted users have less proportions of creation sources from organization invitation and personal projects.

We create new features: one is invited_adoption which indicates that if the invited user is an adopted user; the other is active_span, which is the length from last log_in time to creation time

After cleaning the data, we use Logistic regression and random forest to fit the data. We use features of 'creation_source', 'opted_in_to_mailing_list', 'enabled_for_marketing_drip' and then add features of 'invited_adoption' and 'active_span'.

We find the predict ability is improved significantly after adding 'active_span' features, reaching to 99%. We also check the feature importance after running the models

Logistic Regression

Weight[?]	Feature
+0.460	invited_adoption
+0.044	active_span
-0.000	org_id
-0.015	enabled_for_marketing_drip
-0.081	creation_source
-4.605	<BIAS>

Random Forest

Weight	Feature
0.9625 ± 0.0326	active_span
0.0232 ± 0.0181	org_id
0.0056 ± 0.0105	creation_source
0.0044 ± 0.0063	invited_adoption
0.0027 ± 0.0058	opted_in_to_mailing_list
0.0017 ± 0.0032	enabled_for_marketing_drip

And both ‘active_span’ and ‘invited_adoption’ are important features for predict future user adoption.