Relax Take Home Challenge

The user data and user engagement data were imported into a Jupyter notebook to answer the question, what factors predict future user adoption? The following analysis steps were undertaken.

1. Manipulate dataset to find users who fit the definition of “adopted user”
2. Build a feature set of adopted users by dropping irrelevant columns such as user name and filling N/A values intelligently
3. One hot encode categorical variables to make them numerical
4. Scale the feature set
5. Perform Principal Component Analysis on the feature set
6. Identify which features correlate most with the principal components from the PCA.

The above identified features are the most important factors in predicting future user adoption.

There is some level of arbitrariness in the feature selection process, because all of the features have some level of signal buried in them, however there are a few that actually are the most useful and explain most of the variance in the feature set, and are therefore the most predictive of the target variable. Below is the summary table of the sum of the absolute value of the correlations of each feature with the top 7 principal components, which are the 7 principal components that explain 92% of the variability of the feature set.

org_id	1.792770
SIGNUP_GOOGLE_AUTH	1.705164
last_session_creation_time	1.700941
SIGNUP	1.524752
PERSONAL_PROJECTS	1.438816
ORG_INVITE	1.076612
GUEST_INVITE	1.016688
enabled_for_marketing_drip	0.698910
opted_in_to_mailing_list	0.695712
invited_by_user_id	0.637653

The features in all caps are the one hot encoded features derived from the creation_source feature.

We can see that the most important features are org_id : the group of users they belong to, creation_source : how their account was created, and the last time they logged in, given by last_session_creation_time .