Joseph Petersen - Relax Inc. Challenge Findings

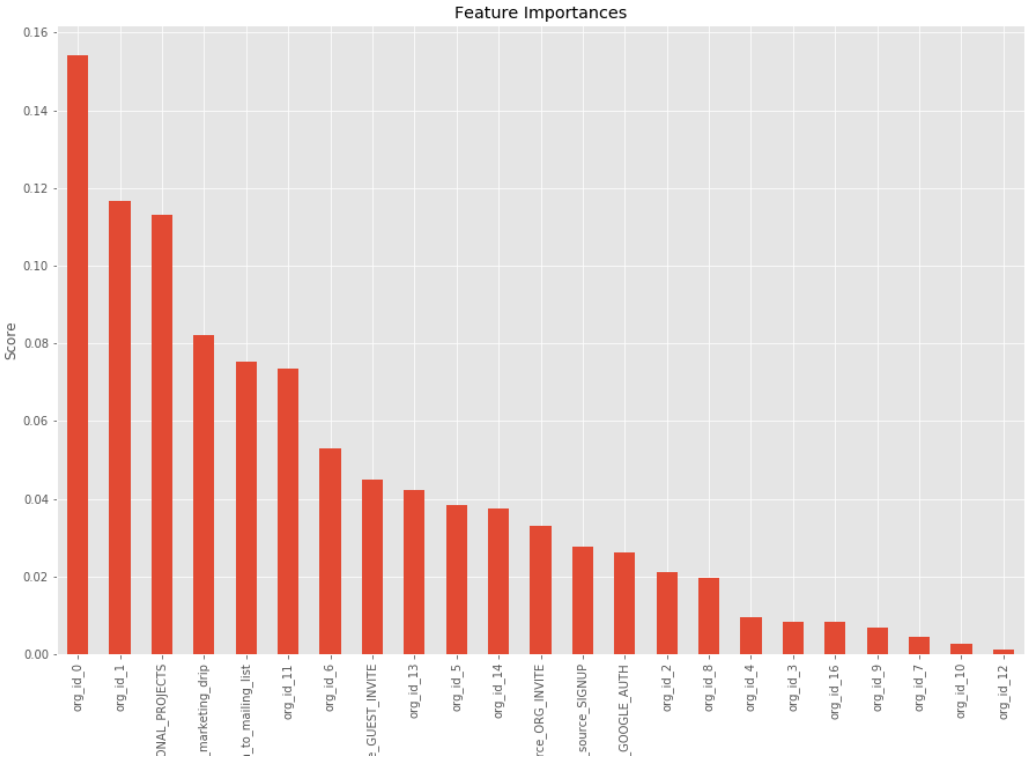

**Figure 1:** The above figure shows the figure importance scores of all the variables. Some variables were categorized and dummied: org_id was categorized down to 17 options based on number of adopted users and source was dummied.

After wrangling the data and creating the response variable column based on the definition of an adopted user being an individual who has logged in at least three times in a seven-day period, the remaining features included: creation_source, opted_in_to_mailing_list, enabled_for_marketing_drop, org_id, and invited_by_user_id. All could easily be turned into dummy or categorical variables for modeling except for org_id which had too large a range to create individual dummy variables. It was suspected that the specific org_id a user belonged to was a good predictor given the uneven distribution of users among the different ones and needed to be kept as valid data. So, it was clustered by number of adopted users and assigned to one of 17 groups. The org_id group assignments are stored in a separate dataframe for reference.

The org_id of the user is the most important predictor. By breaking them into clusters, we can see the large spectrum of feature importance placed on each grouping. Aside from the org_id, whether or not a user created an account by joining someone's personal workspace was a strong predictor. Marketing targeted at individuals with strong personal drive such as freelancers could be effective as well as incentives for bringing in colleagues. If the user opted into marketing emails or has them still enabled also has high feature importance. Further testing on how to encourage new users to enable marketing emails as well as how to retain them on the mailing list would be beneficial in creating adopted users.