Relax D. S. Challenge: Inquiry into Features that Predict User Adoption

By Dovid Burns

To answer the question posed in the relax data challenge, we first had to find a way of identifying which users were “adopted users”. We initially found that from the 12,000 total users in takehome_users.csv, only 8,823 users logged in at all. Digging deeper, we found that only 2,248 of these users logged in on at least three separate days, which is the threshold log in rate of an adopted user. Of this subcategory of users, only 1,656 of them logged in on three separate days within a seven-day period thus making that the total number of adopted users.

We next analyzed potential differences in the features of the adopted users and non-adopted users to look for which factors are important in predicting future user adoption. We found no significant difference in the user adoption rate (U.A.R.) of mailing list opt-in and marketing drip enabled. We did find a significant difference in the U.A.R. per creation year. Users who created their account in 2012 had a U.A.R. of 16.7%, 2013 had U.A.R. of 15.4% and 2014 had a U.A.R. of only 8.9%. Due to the scope of the information available for analysis, it is unclear why the U.A.R. decreased by fifty percent between 2012 and 2014. It is possible that the shorter time frame available for 2014 users to become adopted users is responsible for the lower U.A.R.

We found a statistically significant difference in U.A.R. based on creation source. The highest U.A.R.s were created from guest invite and Signup Google Auth, both above 17%. Org Invite and signup had moderate U.A.R.s with 13.5% and 14.5% respectively. Personal projects had the very lowest U.A.R. of 8.2%. We found a surprising correlation between U.A.R. and the month of user creation. The months with the highest U.A.R.s were June, August and October, with U.A.R.s of 18.2%, 17%, and 17.5%, respectively. The months with the lowest U.A.R. were March, April and May, with 13.9%, 9.8%, and 5.4%, respectively. More information would be needed to explain these statistically significant discrepancies.

Examination of org_id showed there are 417 different orgs these users belong to. This is a huge factor in predicting future user adoption as there are vast differences in U.A.R. by org_id. The org_ids with the highest U.A.R. are 387, 235, and 270 with U.A.R.s of 58.3%, 46.2% and 42.9% respectively. To contrast, there are several org_ids with U.A.R.s of 0%. Lastly, we found a range from 0% - 100% of U.A.R.s by looking at which user invited them. Given the fact that the range of invites by any single user was 1-12, this feature is not statistically helpful in any way as a predictive feature.

In summary, org_id, month of account creation, and creation source all are strong features to predict future user adoption. Year of account creation is statically relevant but needs further investigation if this trend can be used as a prediction feature. Additional data that might be helpful would be the average time per session, the industry of the user, user gender, user job title, and user’s company’s yearly revenue.