Elie Park
August 18, 2022

Relax Inc. Challenge

In the univariate analysis, it was found that the distribution of 'org_id' was right-skewed, meaning that the lower organization IDs indicated larger organization groups. Figure 1 shows that the adoption rate for these groups were higher in the smaller organization groups labelled with higher organization IDs. After converting the creation_time and last_session_creation_time columns into datetimes, the difference between these variables were calculated. Half of the users were active for less than 1 day in total, and the mean time difference was 59 days.

Figure 1. Adoption Rates within Organization ID groups

Figure 2. Adoption Rates within Invitation Groups
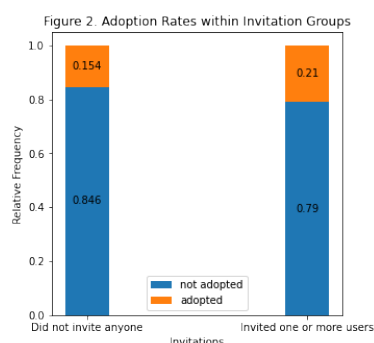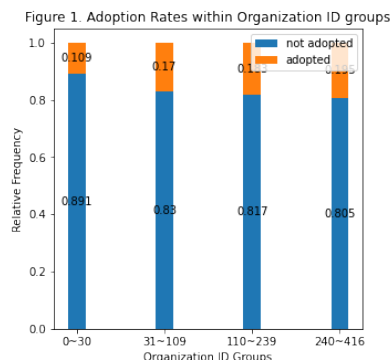
A new 'invitees' column was created telling us how many users were invited by each user. The distribution for 'invitees' was also right-skewed. Those whose accounts were created through an organization as a guest and those who were invited to join another user's personal workspace had the most invitees per user. Users who signed up via the website and using Google authentication had the least invitees per user. Users who did not invite anyone were less likely to be an adopted user.
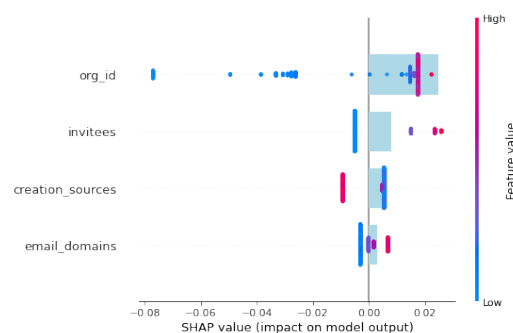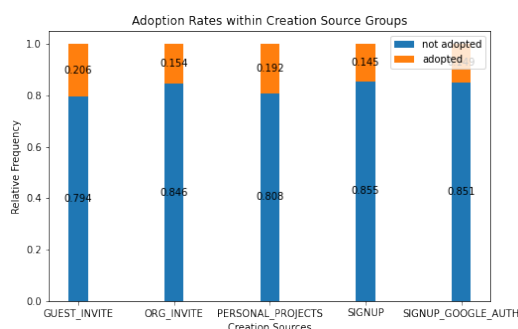
The non-normality of these distributions were accounted for using MinMaxScaler. Since there were 7365 non-adopted users and 1445 adopted users, 1445 counts of non-adopted users were randomly sampled to create a balanced dataset. Out of Random Forest Classifier, Logistic Regression, and XGB Classifier, Random Forest was found to be the model with the highest ROC_AUC score of 0.5855. Organization ID was the most predictive feature, and whether or not the user was enabled for marketing drip or opted into the mailing list was not predictive of adoption. Top three features were congruent with trends mentioned above. Users with random email domains (eg. Not gmail, yahoo, etc..) were more likely to be adopted although there was no correlation found between email domains and organization IDs. There was a slight Cramer's V correlation between email domains and creation source, possibly due to one of the creation sources being google authentication.