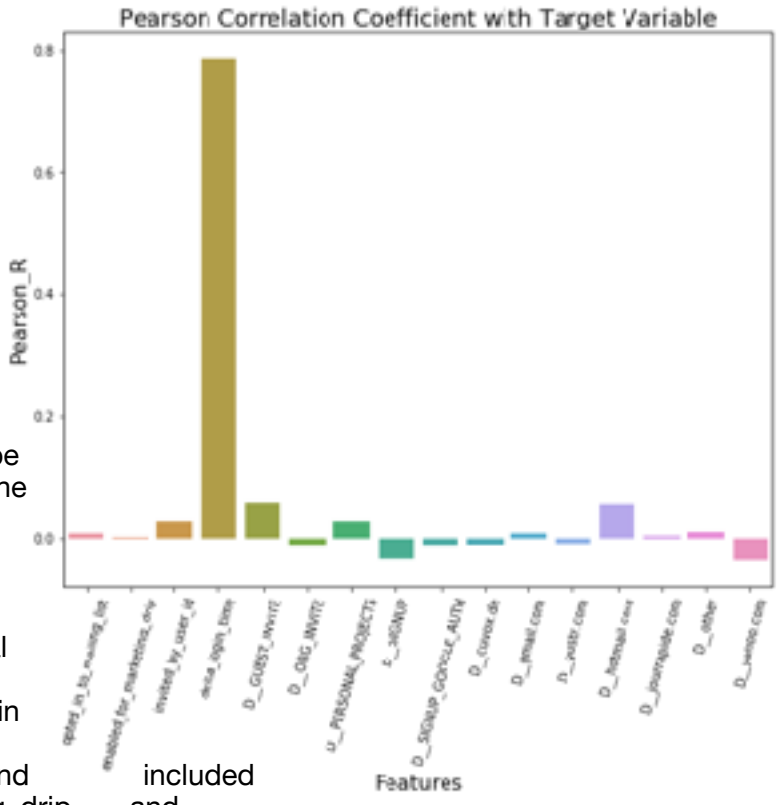
Dan Feldman

## Relax Challenge Results




For this take home challenge, I was presented with two tables, one with user information and another with user login timestamps, and was asked to determine the factors leading to becoming “adopted users.” These are defined as users who logged in three times in the span of a week, and was found to be 18.2% of all the users in the data given. Given the imbalanced dataset, I decided to look at statistical measures of feature relationships between the target variable and the features individually. This was mainly using the Pearson Correlation Coefficient, but given the categorical nature of the target and some of the features, I also used Cramer’s V and the Correlation Ratio in order to characterize these relationships.

Some features were given in the table, and included opted_in_to_mailing_list, enabled_for_marketing_drip, and invited_by_user_id. I created some new features using other columns given such as a delta_login_time (time difference between first and last login session) and dummy variables representing email domains and creation source of the user. I altered the invited_by_user_id variable to be a categorical binary variable of yes or no to deal with missing values and the meaningless quality of the id itself. Given the wealth of email domains used, I only looked at the top six email domains and used a seventh “other” category for the rest.

As seen in the figure, i found that the strongest predictive feature was the delta_login_time, implying that the longer a user is utilizing the service, the more likely they are to become an adoptive user, which makes intuitive sense. Other strong correlations are seen in the D_GUEST_INVITE and D_hotmail.com features, meaning there could be a relationship with those who join the service via an invitation, and those who use hotmail accounts for their emails. The first of those makes some sense to me, while the second is an interesting oddity that would warrant further study. The two strongest anti-correlations are seen with yahoo email account users, and those who signed up with the SIGNUP event. Though not shown on the figure, these correlations all have p values less than 0.05, and are thus statistically significant.

Follow up calculations with Cramer’s V and the Correlation Ratio seemed to echo similar results to the Pearson Correlation Coefficient. However, these relationships only look at the effects of features individually on the target variable, but does not look into any collective effects of the features together on the target. To do that more complex analysis, I would need to do some predictive modeling with machine learning algorithms. Different analyses can showcase different relationships, and so multiple methods would be good for use, such as logistic regression, random forest, and a neural network, to name a few. In order to carry out this analysis, I would also want to look into any possible correlations between the features themselves for colinearity, which could lead to modeling problems. Given that only 18.2% of users are adopted users, I would also want to balance the dataset before using these algorithms, perhaps via resampling.

Given the time constraints of the challenge (1-2 hours recommended), I leave this more complex analysis for future work.