We would like to understand the factors or drivers that lead to login behaviour of users. There are two datasets in CSV format on 12,000 users who signed up for the product in the last two years. The first data set has information in the user like name, email, creation time while the second dataset has information login activity by day. We define an “adopted user” as one how has logged on three separate days on a 7 day period. We would then like to find the factors that predict future user adoption.

To answer our question, we first compute our target variable from the data set “takehome_user_engagement.csv” that has a row for each day the user logged in. The target or the endogenous variable takes the value 1 if she is a “adopted user” by the definition above and 0 otherwise. Clearly, this then becomes a binary classification problem where the dependent variable takes on 2 values, 0 and 1. Next we merge the data on the 12,000 users with the target variable using the unique user id, obj_id. The target variable has 8823 rows, since 3,177 users have never logged in. These users have been given a target value of 0. For our dataset, 1714, or 14.28% of users are “adopted users” while 10, 286 or 85.72% users take the target value of 0.

To predict user adoption, we model the data using supervised learning algorithms like logistic regression, decision trees and random forests. The explanatory variables are creation_source, which takes on 5 values, personal_projetcs, guest_invite, org_invite, signup, signup_google_auth[1], whether or not they are on a regular marketing drip and whether they have opted into receiving regular marketing emails. Since the classes are unbalanced, we balance the classes prior to modeling. This is done so as to avoid a scenario where the algorithm predicts 0 for all users, while achieving an accuracy of 85%. The modeling results, unsurprisingly are not good, with logistic regression we get an accuracy of 54%, decision trees are marginally better at 55% and random forests performing the best, but only slightly, at 56%. For all three models, the test accuracy is slightly higher than the train accuracy, clearly indicating underfitting or bias in the underlying models. This implies that the number of variables or columns in the feature matrix is too small to get any significant results. Thus to improve the results, we would need more data. Some demographic information like age, sex, education, profession, and city of residence on the users will definitely aid in predicting the future user adoption.

[1] We create dummy variables using Python’s one-hot encoder, and drop one of the dummies in our feature matrix to avoid multicollinearity.