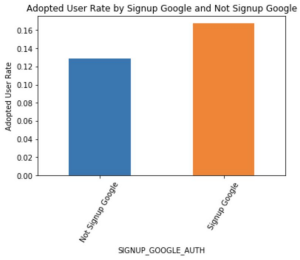
Predicting User Adoption at Relax Inc

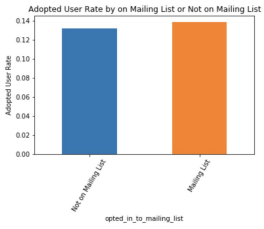
An adopted user is a user who has logged into the product on three separate days in at least one seven-day period. The goal is to identify which features predict a future adopted user given features about the user such as how they signed up for the product, when the user signed and last logged in, which marketing tactics the users is opted, and the invitation source for the user to sign up. Creation source, marketing source, and invitation source effects on user adoption will be explored before the prediction task.

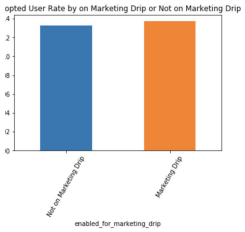
Creation Source: Features that tells us how the account was created.



The two most significant differences in user adoption rate was when the user signed up for a personal project vs when they did not. When the user did not sign up via a personal project user adoption rate was 7% higher. In addition, when the user signed up with a google authorization vs when signed up without a google authorization, user adoption was 4% higher when they had a google authorized account..

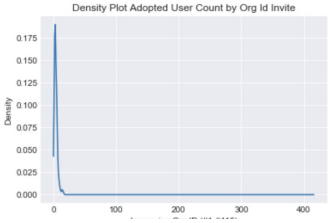
Marketing: Features that tell us which marketing options the user was opted into.



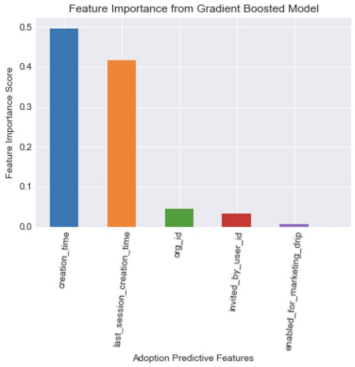


For both the marketing drip and the mailing list the change in adopted user rate was minimal and was less than 1% in both cases. These features do not appear to be strong predictors and most likely will not be top features in the prediction task.

Invitation Source: Features that tell us who invited the user to the product.



The adopted user count is significantly higher in the first 30 orgs than any other subgroup and can be seen in the early spike of the density plot below. Invitation by org id will be used as a feature in the prediction mode because the density model is non uniform and it seems when the org number is less than 30, there are more adopted users.



Prediction Task: The adopted user status was predicted using an extreme gradient boosting tree method. All of the features in the user signup DataFrame were used in the model. A train test split of 0.8/0.2 was used and 5 folds were used in cross validation. The default number of estimators was used (100). **Model Accuracy: 97.4%.**

The top 5 most important features in the model were in order:
1.) Account Creation Time 2.) Last Session Creation Time 3.) Org ID 4.) Invited by User ID 5.) Enabled for Marketing Drip

Although the gradient boosted model does not tell us which values of account creation time and last session creation time are causing adopted user rate, a logistic regression model can be used to find the sign of the coefficient for these features. The sign will give a clue as to how the probability of becoming an adopted user changes as account creation time and last session creation time increase. In the future, if certain days of the week or months want to be checked for their effect on user adoption, the features for each category can be engineered from the datetime values.