Relax Data Science Challenge

The goal of this challenge is to predict “adopted users”, or active users who have logged into the product three separate days in at least one seven-day period. The data consists of the name, user id, how the account was created, creation time, login time, opted in mail list, opted in marketing emails, organization that they belong tom and which user invited them to join.

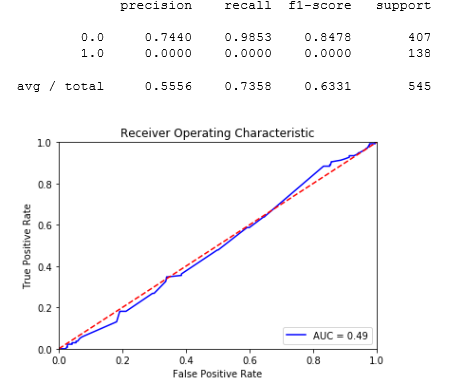
After cleaning through the variables, it was determined that session creation time, invited by user id, name, email, and creation time did not appear to be predictive variables, so they were dropped. In addition, only the first ten organizations were used in the model.

The first alarming observation from observing the data is that there is a class imbalance. As you can see in the below figure, Adopted Users only occur ~17% of the time.



Without making any adjustments to account for the class imbalance, both a gradient boosted classifier and a random forest model was performed. Both had high accuracy statistics as expected, but very poor AUC and precision statistics.

After attempting an undersampling of the larger class(~4,000 records of larger class, ~2000 smaller class), accuracy went down as expected(~87% to ~74%), as well as other statistics from the full sample models. Both the gradient boosted classifier and the random forest models produced roughly the same statistics. Precision went from ~0.85 to ~0.55, recall went from ~0.88 to ~0.74, and AUC remained at ~0.5 for models using the full sample in comparison to the models generated from undersampling.



Most important predictors are opted in to mail list, opted in to marketing, and organization group 4.

Therefore, while additional methods can be attempted to account for the class imbalance (undersampling, back-end scoring metrics, etc), the evidence so far suggests that the variables chosen are generally not predictive enough in generating a model.