Features for Predicting Active Users

The most important features in predicting whether users become active is the time they created their account. However, several other factors are also important, namely: the organization they belong to, if they were part of a marketing drip, and what their creation source was. To find this I first determined how many users fit the “active” criteria of having logged in at least three times in a 1 week period. Of the 12,000 users, only 1602 were active – or about 13%. Then I used a boosted random forest to determine which features best predicted the active users. The model had 92% accuracy, though with tuning I’m sure that could be increased.

With the overall active rate being so low, it seems that further investigation would be warranted into what about the creation time is so important. Did most active users join at certain times? This with the marketing campaign variable this seems to imply that. Which creation sources and which organizations tend to produce the most active users. These would be good to investigate. Additionally, I expected that those who were invited by another user would be more likely to be active, but that feature had no influence (or at least was redundant with another). This could be due to me misinterpreting that information. I assumed that the missing data in that column meant the user was not invited. If that is inaccurate, the additional data could be helpful in making that feature meaningful.

-Matthew Rytting