After running a random forest model with tuned hyperparemeters, overall model performance was not that great, with 85% accuracy, an ROC AUC score of 0.55 and a PR AUC score of 0.16. In my final model I tried removing the object\_id column, or just the I.D. given to a customer when they signed up as I felt this should have little predictive ability and it was deemed as the most significant factor in the random forest model. After dropping this column, the most important features by far were ’org\_id’ and ‘invited\_by\_user\_id’, which stand for the organization the user belongs to and the I.D. of the user they were invited by. I’m still not sure these make a lot of sense as predictive features. Although random forests are scale invariant, meaning you shouldn’t have to scale your variables so that variable importance isn’t skewed by the absolute size of the values, I think this may have been an issue. I tried scaling the variables in the random forest model but there was no affect on the output (as expected).

For future work, I think I could try using some other models that do benefit from scaling variables (such as logistic regression) and I also would implement some resampling or undersampling of the minority class (adopted users) to try and improve predictive performance for adopted and non-adopted users.