Relax Data Challenge Write-up

Sam Fisher

Which factors predict future user adoption?

To explore this question, we used a Random Forest classifier to determine how much useful information each feature adds to task of predicting adoption. Overall, the model was only somewhat predictive, attaining an f1 score 0.62 on a test set of 3960 users (33% of the data). This means that there are a likely a variety of features not included in the dataset that we can use to improve our results. The mailing and marketing drip features, as well as the "personal projects" sign up source turned out to be most relevant. (See bar plot below.)

This is somewhat surprising because the spearman correlation values for the mailing and marketing features are much smaller than their importance assigned by random forest. When modeled with Naïve Bayes, the resulting difference in probability per class closely matches the correlation values between each feature and the class by spearman's rho. While this is a mathematical curiosity (to me), the best course of action is to find more predictive features!