Relax Data Challenge Write-up

Sam Fisher

Which factors predict future user adoption?

To explore this question, we used a binomial Naïve Bayes model to determine the difference in probability between adopting and not adopting given each feature. The model was trained on a balanced set of example from the adopted and non-adopted class. Overall, the model was only a little predictive – attaining an f1 score 0.6 on a test set of 1804 users. This means that there are a likely a variety of features not included in the dataset that we can use to improve our results. Domain and signup source proved to be much more predictive than the mailing and marketing drip features.

In addition to utilizing the included mailing/marketing list features and the signup source, we also used the user's domain name, which proved to have some limited predictive power. The resulting difference in probability per class closely matches the correlation values between each feature and the class by spearman's rho (for additional confirmation).