Relax Challenge Take-home

Nathan Wodarz

2021-07-08

Users who logged in at least three times were likely to become adopted. By definition, all other users could not be considered adopted, as an adopted user must have logged in at least three times. An analysis found little else in the data set correlated to whether users became adopted.

We examined a total of eight features in determining whether we could predict a user became an adopted user. These included three features that were included in the original data set: `opted_in_to_mailing_list`, `enabled_for_marketing_drip`, and `creation_source`. The other features were derived from the ones given. The feature `session_count` gave the number of distinct logins for the user, `invite_count` the number of other users invited by the specified user, `was_invited` indicated whether the user joined via invitation, `org_count` was the total number of users who share an `org_id` with the specifed user (including the user themselves), and `creation_days_since_start` was the number of days that the user created their account (taking the earliest account as day zero). The `creation_source` feature was one-hot encoded into five individual features, giving a total of twelve predictive features for modeling purposes. Other features were ignored.

Very little correlation was seen between features. Users who had opted into the mailing list were more likely to be enabled for the regular marketing drip. Users who joined later (predictably) had fewer sessions and had invited fewer people. The only other correlations of potential note were that whether the user was invited correlated with the `creation_source` encoding, as two of the encodings referred specifically to invitations from organizations (`GUEST_INVITE` and `ORG_INVITE`). The target feature `is_adopted` had very little correlation to the other features.

The primary difficulty arose from the fact that the data set is somewhat imbalanced. Of the 12000 total users, only 8823 logged in at all, and only 1602 of those became adopted users. Even when restricting to the 2248 users who logged in at least three times, there is little correlation to other factors.

A Principal Component Analysis was performed, but was found to only reduce two of the thirteen total features. Our remaining time was spent trying a variety of classification models. Logistic Regression, K-Nearest Neighbors, Ridge Regression, Naive Bayes, Random Forests, Support Vector Classifiers, and Gradient Boosting were all used. Most of those predicted all or nearly all of the users to not be adopted. When using balanced class weights, we saw some improvement but no method was noticeably better than the baseline.

Several methods went untested, particularly using methods designed for imbalanced data sets. No under- or over-sampling was performed. Evaluating these techniques could allow progress to be made. Additionally, it may help to have additional data. In particular, the length of sessions and additional information about particular organizations may be of use. Since users who logged in at least three times were likely to become adopted (1602 out of 2248), it might be worthwhile to use the number of sessions as the target variable for regression analysis.