

Relax Challenge

Dylan Neal

Relax Inc. is a company with a digital user-based product. Two data sets were provided: one containing features related to how each of 12,000 users signed up for the product, and one containing user timestamps for each time they engaged with the product over a two year timespan. An “adopted” user is one who interacted with the product on at least 3 different days over the span of a single week.

After cleaning the data it is clear that they are imbalanced, with only 1 out of every 7 users end up classified as “adopted”. It is in the best interest of stakeholders to better identify features associated with eventual user adoption. Recall is selected as the appropriate scoring metric, as identifying the true positives of the minority class (adopted users) is essential. If there is an opportunity to preemptively upsell users that are flagged as future adopters, that would be advantageous. More is lost by not engaging with future adopters than by engaging with those that are less likely to adopt, due to the class imbalance.

After generating the class labels, unimportant features were dropped (including name, object_id, creation_time, and last_session_creation_time). User email domains were parsed from the email feature. The top 7 most frequent email domains remained, while the remaining 1,177 domains were combined into an “other” classification. The same practice was done for org_id, reducing the number of unique org_ids from 417 to 11 total. Later modeling would show that inclusion of the org_id feature did not improve modeling, so it was dropped entirely for the final model.

The data were scrubbed and modeled with a Decision Tree Classifier with class weights set to 7:1. Final recall score for the adopted class was 0.55, which is not horrible with a 1:7 class imbalance. Feature importance was extracted (see below for chart). The most important feature is creation_source_PERSONAL_PROJECTS, which shows that users creating their accounts for personal projects are less likely to eventual adopt the platform than those joining for other reasons (such as Google authentication or organization invites). It also showed that users which enable marketing drip and opt for mailing lists are more likely to end up classified as adopted. A/B tests could be run in the future to determine how much impact mailing lists and marketing drips actually do have on user adoption.

