Future User Adoption

Corey Wade

October, 2018

In tables provided by the RDC, it was essential to first build an "adopted_user" metric by searching timestamps for 3 days of use within a 7-day period.

It was necessary to eliminate the time columns due to confounding variables. Since adopted_users are determined directly from time, it's possible to obtain a meaningless accuracy score of 1.0.

Of the remaining categories, "creation_source," "opted_in_to_mailing_list," "enabled_for_marketing_drip," "invited_by_user_id," and "org_id," creation_source needed to be factored into integers because it consists of strings.

The correlation coefficient revealed that org_id was the most influential factor at 7%.

Machine learning tests initially revealed a class imbalance. Many respectable ML classifiers, like Logistic Regression, predicted no adopted users. With 14% adopted users overall, resampling was necessary to balance the classes.

After resampling with cross-validation and grid searches, Random Forests performed best at 78% accuracy, followed by Decision Trees and K-Nearest Neighbors.

These methods can be used in conjunction with Recursive Feature Elimination, and "feature_importances_" to determine the most influential features.

According to feature_importances__ on multiple classifiers, org_id was by far the the most influential factor in the 80th to 90th percentiles. Creation_source stood out as a clear number two.

Digging deeper into creation_source, I drew up the following tables.

ALL USERS

ORG_INVITE	0.354500
GUEST_INVITE	0.180250
PERSONAL_PROJECTS	0.175917
SIGNUP	0.173917
SIGNUP_GOOGLE_AUTH	0.115417

ADOPTED USERS

ORG_INVITE	0.346618
GUEST_INVITE	0.222826
SIGNUP	0.182367
SIGNUP_GOOGLE_AUTH	0.144324
PERSONAL_PROJECTS	0.103865

Further analysis is warranted in order to dig deeper into org_id. This column and the creation_source column can be one-hot encoded into identity matrices to reveal more nuanced results. Collecting more data, and running time series analyses are also advised.