Relax, Inc. Take-home Challenge Report

Nils Madsen, 11/9/18

Approach

The approach taken for this challenge was to leverage the capability of machine learning algorithms to estimate the importance of features for predicting the target. A random forest model was fit to the data, and its estimation of feature importance for predicting user adoption was used to guide an exploratory analysis of the features.

Data Cleaning and Feature Engineering

Certain verifications were performed on the data, such as confirming that all the user IDs were unique, and that all user IDs in the login data were in the set of users in the user data.

Users who created an account in the last two months of the data period (April/May 2014) have a strong drop-off in adoption rate due to an artifact in the dataset, whereby there wasn't login data far enough into the future for these users to have a good chance of being considered adopted. These users were dropped from the data.

Some features of the dataset were removed, such as name (not useful) and last login time (target leakage) from user data, and the visited field (only one unique value) from the login data. New features were extracted, such as the target variable, datetime features (month, day, time of day, etc.) from account creation time, whether the inviter was considered adopted at time of invitation, and the user's email provider.

Feature Importance and Exploration

For random forest training, the users were split into two groups based on when they created their account (first and second year of data period). The model was cross validated on these two groups to verify its ability to predict the target from the dataset, and it achieved ~84% accuracy.

Random forest ranked account creation time as the most important feature, even after removal of the last two months of users. Investigating trends across months, days of the week, and time of day did not reveal obvious patterns. The model may be detecting certain high-adoption times throughout the year.

The user's organization was the next most important field, indicating that some organizations are more likely to convert users to adopted status, perhaps because they use Relax's service as part of their day-to-day operation. The model also recognized email provider and account creation source as valuable features; hotmail and gmail users were most likely to adopt the service, as were those who were invited as guests to an organization, and those that signed up using Google authentication. The model did not see much value in whether the user was on the mailing list or marketing drip.

Future Directions

This is a rich dataset with a lot of potential for feature engineering and investigation into how the features are predictive of user adoption. One of the most promising avenues for further research is investigating exactly how user account creation time is predictive of user adoption.

It would also be useful to have data on user logins for at least 2 months after the end of the timespan for user account creations. This would allow for the inclusion of users who created accounts in April and May of 2014.