Summary - Example Take-Home Challenge: Relax Inc.

The goal of this challenge was to identify 'adopted users' of a productivity and project management software platform created by Relax Inc. and to identify which factors predict future user adoption. Two files were provided. The first file 'takehome_users.csv' is a table with information related to 12,000 users of the platform who signed up in the last two years and includes 9 different attributes in addition to their unique identification in the system. The second file 'takehome_user_engagement.csv' is a table that includes a row for each day that a user logged into the product.

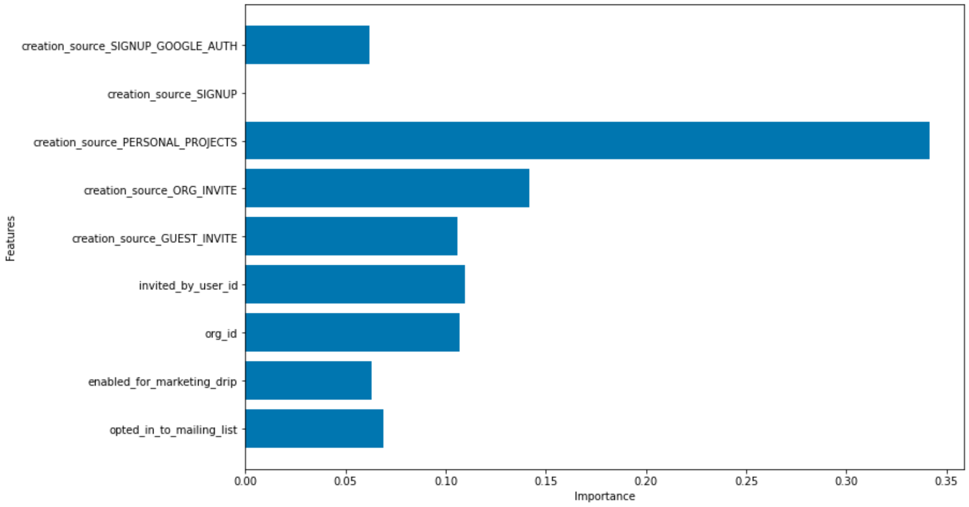
Two techniques were used to identify the adopted users from the time stamps. The second technique (a pandas rolling function followed by a groupby max) provided the expected results and was therefore merged with the first file to become the target variable (y) in the modelling step. Since only a total of 8823 users were included in the second file, and there are 12000 users in the first file, it is clear that most users actually fall into a 3rd category that never logged in at all during the period recorded in the second file. If more time were available, I would recommend assigning a different value to these users and using a multi-classification approach, but given the time constraints of 1-2 hours, I opted to simply assign those users a value of 0 as well in the 'adopted user' column.

During the EDA step, it was found that all 5 types of account creation sources have some adopted users, but the lowest percentage of adoption is from the 'Personal Project' category. There was very little difference in means between the adopted and non-adopted users as to, whether they were opted into emails, or signed up for marketing drips. It was also noted that certain organizations and certain users had a higher percentage of adoptees. Additional features could have been created from the user names and email addresses (such as gender and email domain), but in the interest of time this was not attempted. Since user login data was used to identify adopted users, neither the account creation date or last login date or any other features derived from them were used as features in the model.

An XGBoost classifier was used because it runs quickly and works optimally when the number of features is less than the number of observations in the training data. It performs well when data has mixed numerical and categorical features or just numeric features and it is easy to extract the feature importance.

Results

The feature importance results confirmed the earlier observation that creation sources are all important, particularly those that are personal projects. There appears to be a high inverse correlation between adopted users and those users from personal projects. Other factors were less important for prediction.


creation_source_SIGNUP_GOOGLE_AUTH
creation_source_SIGNUP
creation_source_PERSONAL_PROJECTS
creation_source_ORG_INVITE
creation_source_GUEST_INVITE
invited_by_user_id
org_id
enabled_for_marketing_drip
opted_in_to_mailing_list
Features
Importance