Data loading and exploring

Loaded both csv files into a pandas dataframe with the ISO-8859-1 encoding. The user_engagement file had no missing data, while the users file had 74% values missing for last_session_creation_time, and 53% values missing for invited_by_user_id.

takehome_user_engagement data

Converted the timestamp to a datetime_index so that datetime operations could be performed. Sampled data in one-week periods and grouped by user_id. A user was active if he had logged in at least thrice, each time on a separate day of the week. Grouped user_ids per week. If a user had visited three of more times, changed that value to 1 else saved it as 0. Grouped by user_ids and aggregated the total number of visits for each user_id. Added a new column, adopted_user for the target variable. Populated this column with a 1 for three or more visits, and a 0 otherwise. Dropped the visited column as it was not needed anymore. Renamed the user_id column to object_id to make it consistent with the column in the users dataframe. Merged the wrangled data with the takehome_users data.

Merged Data

Dropped features that did not seem to have a significant impact on predicting active users. More than 50% of the users were not active/adopted users, and more than 50% had opted out of the mailing list and the marketing drip. There were too many invited_by_users, with a lot of missing data. This feature would not have helped with analysis, so dropped it. There are 1147 org_ids, but the number of users who signed up through org_invite was higher than the other creation sources. So, this seemed to be an important feature in predicting active users.

Model Building and Predicting

This is a Classification problem as it has two binary outcomes.

One-hot encoded the categorical variables. Split the data into train and test, fit the training data to a Random Forest Classifier, and made predictions on the test data.

The model predicted creation_source as an important predictor, with org_invite being the most important amongst the other creation sources.

Conclusion

The model gave an f1-score of 80%. Since dimensionality reduction results in better modeling, grouping the org_id before building the model would have given better results.