Relax Take Home Challenge

The data came from two csv files, one containing user login information and the other containing user information. The task was to identify which factors predicted future user adoption.

Data Wrangling

I first resampled the dataframe containing the login information into weekly counts then kept only the users that had logged in over 3 times per week (labeled as 'active'). I then executed an inner merge with the dataframe containing the users information therefore ensuring that only active users were left. I then used this dataframe to create a new column on the original users dataframe to indicate whether users were active or not. I then decided to focus only on six columns for the final modeling phase : 'creation_source', 'opted_in_to_mailing_list', 'enabled_for_marketing_drip', 'org_id', 'invited_by_user_id' and 'active' columns.

The wrangled dataframe had labels in the 'active' column with 0 representing an inactive user while 1 represents an active user. This column was heavily imbalanced with the majority of users marked 0 (inactive). The features which had a high correlation with the 'active' column are org_id_5, org_id_6, creation_source_GUEST_INVITE, org_id_12, and creation_source_SIGNUP_GOOGLE_AUTH. Of these five features, the most intuitive ones are creation_source_GUEST_INVITE and creation_source_SIGNUP_GOOGLE_AUTH. The first one makes sense because most people have a higher likelihood of using a service recommended by a person they trust. The second one also makes sense due to the ease of logging in with one's Google credentials, therefore negating the need to setup new passwords. Organizations 5, 6, and 12 might require their employees to use the service which might explain the high usage levels.

Predictive Modeling

Because the objective is to classify whether a user will remain active or not, I ran the data through a Random Forest Classifier and Gradient Boosting Classifier. Both performed similarly with an accuracy of approximately 88% and F1 score of 47%. This reflects the imbalance in the data with the end result being models that perform better on true positives versus true negatives.

The most important features for the Gradient Boosting Classifier were creation_source_PERSONAL_PROJECTS, org_id_0, org_id_5, org_id_11 and org_id_6. As mentioned previously, the organizations might require their employees to use the service which could drive increased retention rates. Users who interact with the service for a personal project might also have low attrition rates because they're motivated by their project.

Summary

Despite the limitations of the data, it is clear that affiliations with certain organizations mean less attrition. It is worth finding out the characteristics of org_id_5, org_id_6 that make their group members more likely to continue using the service. More actionable steps for the company would be to encourage existing users to recommend the service to their networks perhaps by creating incentives.