Relax Inc. Challenge

Problem Statement

I am given two datasets and asked to identify the factors that predict future user adoption. User adoption is defined thusly: " . . . as a user who has logged into the product on three separate days in at least one seven-day period."

Data Wrangling

I imported both datasets as pandas data frames and inner merged them. I established a new index for the data frame, getting rid of the now irrelevant "object_id" column. Next, I checked for any null values. The column "invited_by_user_id" had over 90,000 null values that I did not have a convenient way of imputing for, so I dropped all null values. Lastly, I checked for duplicates and printed the head of the data frame, assuring that the changes made were successfully implemented.

Feature Engineering

In this section, my objective was to capture the meaning of user adoption in a column. To do so, I converted the column "time_stamp" into a datetime object using pandas' "to_timedelta" function to take into account a 7-day period. Next, I created a "wk" column, standing for a week. With these two procedures implemented, I went on to assign an object "group", which is grouped by both columns "wk" and "user_id" and counted by the column "time_stamp". Now, I'm able to create a column "adopted" based on the object "group". Now, we have a column that captures the definition of user adoption per the problem statement. Next, I think to see the distribution of the "adopted" column; it is clear that there isn't a class imbalance.

Preprocessing

Given that the objective of the analysis is to find the factors that most predict user adoption, I chose to set up my workplace for classification. First, I dropped several unusable variables. Second, I binarized the "adopted" column. Third, I one-hot encoded the object column "creation_source". And lastly, I split the data using sklearn's train_test_split function.

Modeling

I decided to implement a RandomForestClassifier model. I instantiated, fitted, and predicted using the model, followed by assessing its performance via three methods: (1) attaining its accuracy score, (2) getting a classification report, and (3) getting a confusion matrix. The model performed very well, with a recall score of 99%. Lastly, I listed and graphed the features that were most important in the model. The features with the greatest predictive power for user adoption were 0, 3, and 4. These stand for, respectively: last_session_creation_time, org_id, and invited_by_user_id.

Takeaway

It seems that taking into account the time of a last session (its creation time), an organization's ID, and the ID of the user who invited a particular user are all useful features for predicting user adoption.