Relax Challenge - Report

The first step in the analysis was to calculate which users were adopted. To accomplish this, for each user I created an ordered list of timestamps. Using this list I could calculate the number of days between every set of three consecutive visits that the user made. If one of these differences was less than or equal to seven, we can label the user as adopted.

After this, there were some null values to be dealt with. The last_session_created_time had null values, but it was decided that this variable would not be used in any analysis due to its temporal nature and the fact that the frequency of logins would be highly correlated with our target variable and may skew results. The null values in invited_by_user_id were dealt with by converting it to a Boolean variable where non-null values were True and null values were False.

Dummy variables were also created for the categorical variables in the dataset.

At this point, all of the feature in our dataset were Boolean in nature, so bar plots were done to visualize the data and get an idea of proportions of True/False values among the user's adopted status. The bar plots were all similar in nature and did not show much difference between the adopted vs. non-adopted groups. The plots did show the unbalanced nature of the data (far more non-adopted than adopted users).

From here, a simple random forest model was fit to get an idea of the important features. Out of the box, the model achieved a good accuracy score of about 86%. The features that were listed as most important were related to Creation Source.

Of course, some model tuning could be done or other models tried to try to increase accuracy. There could also be more data munging done to generate new features that might be useful. However, I suspect the relatively high out-of-the-box accuracy is due to the unbalanced dataset and the model predicting non-adoption more often than adoption.

I feel the original dataset is largely uninteresting with many columns such as name, email and creation data that are not of much use for analysis. Some more detailed demographic data about users might bring more insight, or better information regarding time spent using the product or which features of the product they used most might provide a basis for a more interesting model.