Initial Data Analysis and Modeling

The dataset contained 12,000 new users over a period of two years. From this data set, only 1500 users could be considered as adopted users and over half of the users created an account but never logged in for the first time. Many of the email domains appeared to be fake domains, which implies that there was no email filter for removing fake email addresses. The supplemental dataset revealed that no user logged in multiple times per day. The dataset overall lacks a lot of information about the users considering that there is only one column that gives a little piece of information about them from the supplemental dataset.

The approach to solving this problem was to create a column called 'adopted_user' by resampling the dataset to determine the number of users that had three or more logins in a period of seven days. An additional column, 'daily_visits', was extracted when merging the two datasets together. Null values were filled in based on what made the most sense. The binary and numerical values of adopted and unadopted users were plotted against each other, with the daily visits having any visible differences. Two columns, 'creation_to_active' and 'login_perc', were created to complement the number of daily visits. Plots of these two variables show very strong differences between the target and non-target groups, as shown in Figure 1.

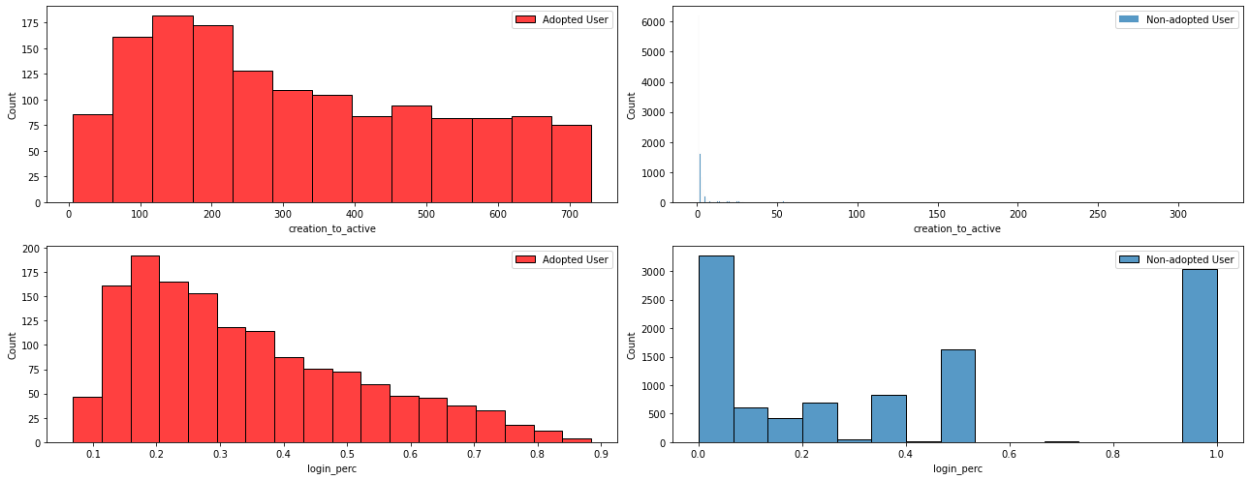

Figure 1 - Histogram plots of creation_to_active and login_perc for both adopted and unadopted users.

The data was preprocessed in two ways. The outliers, which were processed using StandardScaler from the preprocessing library, were very frequent and it was irrational to have a different group of outliers converted to the same value. The categorical values were replaced with dummies. Feature importance was created using the Random Forest Classifier, as shown in Figure 2.

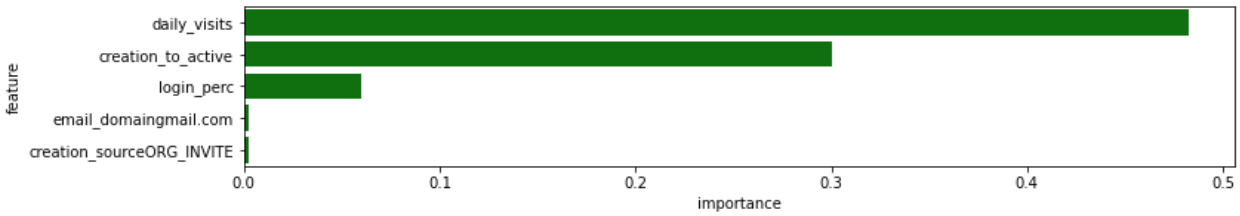

Figure 2 - Feature importance of the merged user's dataset using Random Forest.

The results show that three variables have the most impact on the problem, daily_visits, creation_to_active (the amount of time from account creation to the account being active), and login_perc (derived column which is the percentage of daily_visits versus creation_to_active).