From the 12,000 new users over the span of two-years, only about 1500 users, or 12%, could be considered as adopted-users. Slightly more than half of the total users made an account but have no records of ever logging on. Additionally, it does not seem that there is an email filter as many of the email domains appear to be fake. According to the supplemental dataset, there was no history of a user logging in multiple times per day. Overall, the dataset itself is lacking a lot of supplemental information about the users as there is only one provided numerical column (self-constructed from supplemental dataset).

My approach to the problem was to first define the target column "adopted_user" by resampling and grouping the supplemental dataset to determine the users that have history of 3 login-days within the span of 7 days. During that resampling period, I decided to extract an additional column, daily_visits, to load when merging.

I filled in the null values according to what I believed made the most intuitive sense. For visualization, I plotted the binary and numerical variables of the adopted and non-adopted users against each other, but there were only visible differences in the daily visits. As a result, I engineered two variables to compliment the daily visits: "creation_to_active" and "login_perc". Visual representations of these two new variables show strong differences between the target and non-target groups (Figure 1).

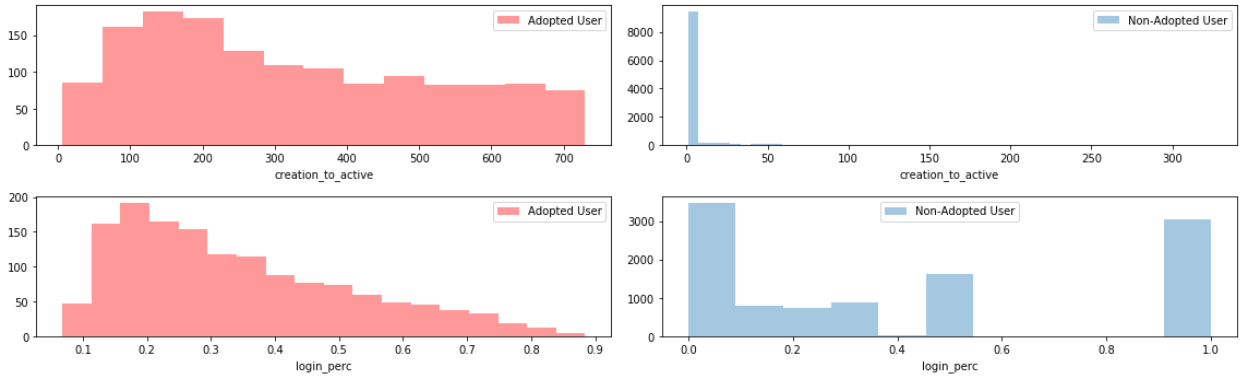

Figure 1. Histogram of creation_to_active and login_perc columns for both the adopted users and non-adopted users.

Instead of dealing with the outliers manually, I decided to use StandardScaler from sklearn.preprocessing. My rationale for doing so is that the "outliers" were so frequent in occurrence that I felt it was irrational to have the different spectrum of outliers replaced with the same value. For the categorical variables, I constructed dummy variables for each of the categorical variables. With the preprocessed dataframe, I calculated feature importance using the RandomForestClassifier (Figure 2). Results suggest that the 3 numerical columns have the most impact.

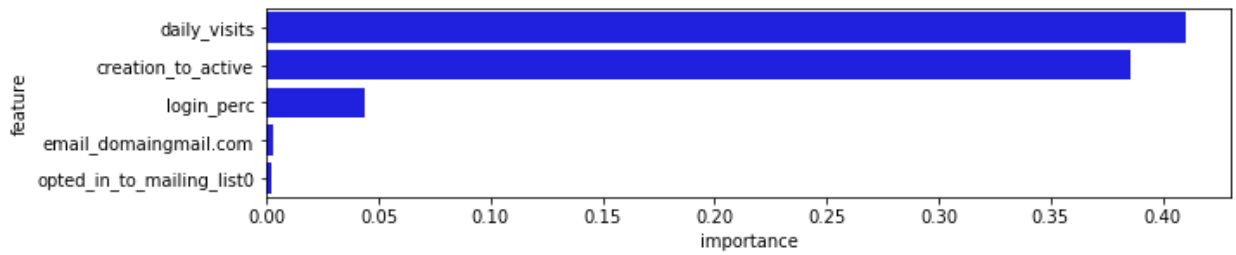

Figure 2. Feature importance of the merged users dataset using RandomForestClassifier.