Relax Challenge Report

*Data Wrangling*

There are two datasets in this project. One is the users’ information dataset that contains 12000 entries with 10 columns. The other dataset contains users’ engagement information that has 207917 entries and 3 columns. When I first loaded the datasets, there are missing values in the columns “last\_session\_creation\_time” and “invited\_by\_user\_id” from the users’ information dataset. For the “last\_session\_creation\_time” column, I filled the empty cell with the mean value of the column. For the “invited\_by\_user” column, after discovered that 0 is not one of the inviter’s id, I filled the null values with 0. After filling in the null values, I changed the “object\_id” from users’ information dataset to “user\_id” to merge with the users’ engagement dataset. The merged dataset contains 207917 entries and 13 columns. Finally, I converted the time-like objects to DateTime object for better operations.

*Identify Adopted Users*

To identify adopted users, I created a column named “wk” that access the week number of “time\_stamp” (timestamp of each time a user login). I grouped the dataset by “wk” and “user\_id” and counted the frequency of the week number by each user. I dropped the week number and merged the frequency count dataset with the original dataset. I created an “adopted user” column that has a value of 1 if the frequency count of the week number is greater or equals to 3, if not, the user received a 0 for the row. The dataset with “adopted user” has 207220 entries and 15 columns.

*Identify Factors Predict User Adoption*

To identify factors that predict user adoption, I need to find the features’ importance through the Decision Tree model. I first dropped memory heavy, unnecessary, and non-integer columns because they will increase the complexity of the model. Then I identified the target feature “adopted user” and non-target features. I created dummy variables for columns that contain the category variable. I split the data into training and testing sets (in case predictive models are applied and need evaluation). I fitted the training set into the Decision Tree model and created a feature importance data frame that excludes the features contributed 0 to the model. I plotted the data frame and the resulted plot is the following: