REPORT ON RELAX CHALLENGE

After hours spent on working on these two datasets I came to a conclusion that the features we have are not enough to make an accurate predictions.

First I cleaned up both of the datasets, I changed the NaN values to 0 for all float columns we had in user_data. I also converted the columnS with data to date time columns.

I kept all boolean columns as 0 and 1.
I resampled engage_data so that I would get weekly visit numbers for each user_id. I later created a 'active' column for user_data data frame (renamed as user_merged) based on the info on engage_data. Because I found the data on user_data insufficient, I also calculated the average weekly visit number for all all users recorded on engage_data so that it could be an important column to predict the 'activeness' of a user. This made a lot of sense since the company determined the activeness of a user based on the number of visits they had per week.

However, when I added that column to user_merged data frame, it was clear that other features seemed much weaker to predict 'activeness' of a user. Heatmap visualizations and as well as feature_importance number of RandomForest Classifier made it clear. Perhaps I made a mistake and miscalculated the feature importances because of a loss of data quality. Nevertheless, I'd advocate that the data quality was low at the first place.

I would love to further discuss the data with the client before moving onto the creating a machine learning function to predict a certain user's 'activeness.'

Thank you