Relax Challenge
Lakpa Sherpa

First, I loaded data from takehome_user_engagement.csv file which had each user's usage summary as a pandas dataframe df_user_engagement. There was some issue loading data from takehome_users.csv, so I converted that csv file to takehome_user1.xls file. Then, I loaded the file as a dataframe df_user. I converted the time_stamp data from df_user_engagement from string to datetime. I grouped the df_user_engagement by user_id and weekly time_stamp with summed number of visit. I saved it to a new dataframe named df_weekly_engagement. I created a new column named adopted_user which saved integer 1 if user visited at least 3 times in a week and 0 if use did not visit at least 3 times in a week. Then, I created a new_dataframe called df_adopted_user which saved user_id of all the adopted_user only. After that, I joined df_user and df_adopted_user using left join.

The new dataframe df have missing or null data in last_session_creation_time, invited_by_user_id, user_id, and adopted_user columns. We have object_id which is user_id, so we can remove the column user_id because it has missing values. We will fill null values with 0 in adopted_user because those users who are not adopted_user should be 0 or False as adopted user. invited_by_user_id null values can also be set to 0 because the user_id starts with 1, and if no other user invited this user to signup then 0 means no. For last_session_creation_time null values we need to discuss with our team or manager. In this project, I will not include name, object_id, email_address, creation_time, last_session_creation_time, and user_id. I will only include creation_source, opted_in_to_mailing_list, enabled_for_marketing_drip, org_id, invited_by_user_id as features and adopted_user as label. I will transform invited_by_user_id to a boolean 1 for True and 0 for False and take care of missing values.

In my opinion and looking at the data, the factors important to predict future user adoption would be creation_source, opted_in_to_mailing_list, enabled_for_marketing_drip, org_id, and invited_by_user_id. Then, I conducted a chi-squared test to check the dependency of the features with adopted_user. The Chi-squared Test suggest that creation_source and invited_by_user_id are the important factor which predict future user adoption. There is a class imbalance issue which has to be taken cared before we build any predictive model.

Here is the link to the codes and graph.
https://github.com/sherpalakpa1/relax_challenge/blob/master/relax_challenge.ipynb