Relax Inc. Take-Home Challenge

1) To define an ‘adopted user’, I did the following steps:

- The two dataset 'takehome_users.csv', and 'takehome_user_engagement.csv' were imported as df1, and df2 respectively;
- The 'time_stamp' column in df2 was converted to datatime format;
- df2 was sorted by 'user_id', and 'time_stamp' for subsequent analyses;
- For each user, the sum of visited times on any 7-day window was calculated; if the sum of visited times on any 7-day window for each user was equal or greater than 3, then it returns True, else it returns False;
- Then, I aggregated by each user, the sum of the above True/False values. If there was at least one 'True' value for each user (i.e. a user has logged into the product on 3 separate days in at least one 7-day period), then assign value 1 to the new variable 'adopted_user', else assign value 0;
- Eventually, I created a new dataframe – df3, which has two columns: ‘user_id’, ‘adopted_user’

2) In order to identify which factors predict future user adoption, I first of all, applied a left join on df1 and df3, based on the unique user id. This create a new data frame – df. Then I did the following pre-processing to prepare data for logistic regression and random forest to output variable importance:

- Filled NAs in df[‘adopted_user’] with 0, assuming if a user did not have login information, then it can be counted as not a adopted user; similarly I filled NAs in df[‘last_session_creation_time’] with 0.
- I created a new variable ‘invited_by_user’ , which was assigned 1 for those that were invited by another user, else assigned 0. I also created a new variable ‘org_id_r’, which assigned all the groups that have a percentage less than 0.5% as ‘others’.
- Select predictor variables including ‘creation_source’,’last_session_creation_time’,’opted_in_to_mailing_list’,’enabled_for_marketing_drip’,’invited_by_user’, ‘org_id_r’; and select ‘adopted_user’ as response variable
- Create dummy features for categorical variables; and applied a standard scalar to predictor variables

As random forest performed better than logistic regression model. I used the results based on the permutation importance analyses of random forest, which suggested that some important factors that predict future user adoption include 'last_session_creation_time', and 'invited_by_user'. Those with more recent last login time, and those invited by an existing user are more likely to be adopted users.