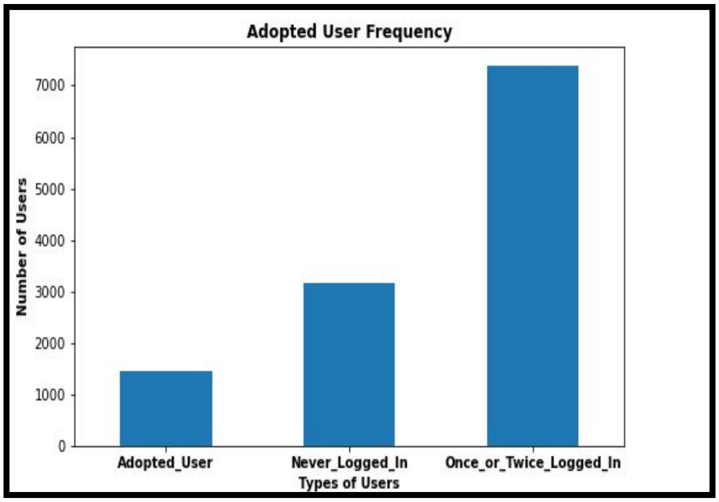
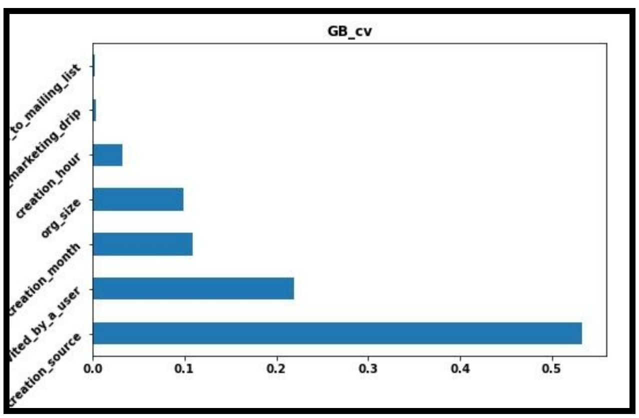
Take Home Challenge: Fariha Baloch



Objective: Identify features that predict future Adopted_User.
Approach: The end goal here is to predict a user's adoption level which I classified into three levels, Adopted_User, Once_or_Twice_Logged_In and Never_Logged_In. The distribution of users per adoption level is shown on the right, which also shows that the data is quite unbalanced.
Feature Selection: In order to train a model I choose the features that a user will most likely be able to put into his/her creation of an account phase or fields that are automatically filled in like creation_dates. Added year, month and day to each user's record from creation_date; Added org_size feature by looking at the total data present from different organizations.
Model: To predict type to a user several classification models were tested. Gradient Boost showed best accuracy. Following shows which features GB gave importance to.The accuracy of this model is 50%.



The top 3 features of importance are creation_source, invited_by_a_user, and creation_month.
Consider a scenario where the marketing team wants to push out new deals to the users in hopes to convert users to "adopted users". Their primary target should be 'Once_or_Twice_Logged_In" users. So if the model predicts a user as that and if the prediction is wrong, there won't be any huge loss. Therefore since Recall score (0.51) for this model is less than Precision (0.63) for class 1 (once or twice logged in) it is a 'not so bad" model to adopt.
Based on the figures below class 1 users can be targeted in May(5) and in 'ORG_INVITE' category to get more Adopted Users.
Other data that could have added value:
User's Age and User's Job Title