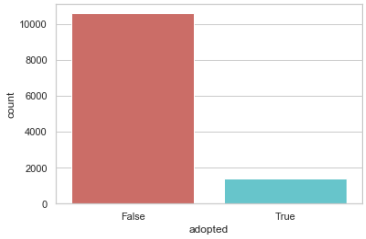
Asana Assessment for Data Science New Grad Role

Chintana Lakshmanan(chintanalakshmanan89@gmail.com)

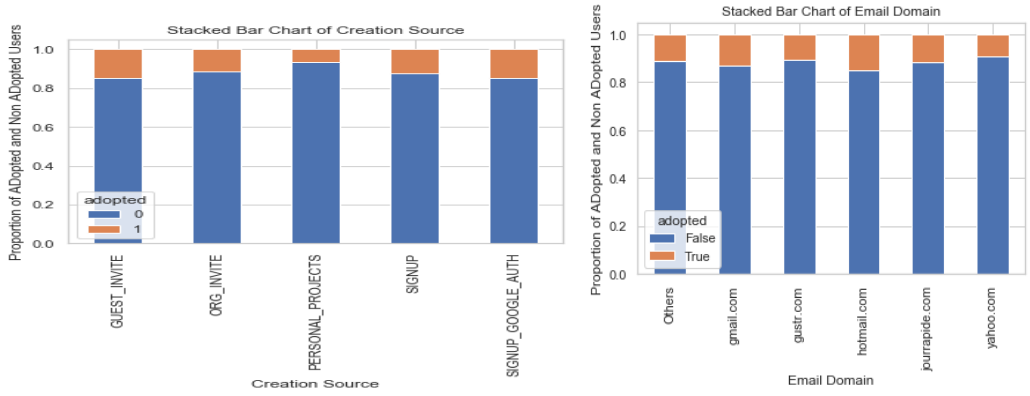
After calculating the Adopted users as per the definition provided we identify that there are 1403 adopted users from the total 12000 users as in the below graph



We then look at factors to predict the Adopted Users. For the sake of this analysis for prediction we remove some variables such as object_id, name ,email etc . We then pre-process the rest of the columns to make it in a more suitable format for analysis and modelling

Below are the columns we ultimately use for analysis

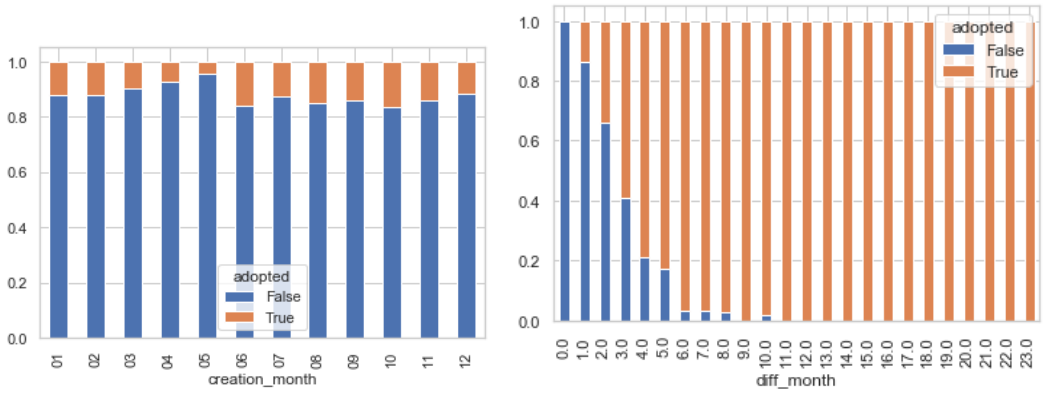
1.) **The creation source. (plot bottom left)** The orange indicates Adopted Users and the blue indicates the 'Non-Adopted' user. We see a good difference in the proportion of Adopted and Non-Adopted users for each creation source and hence could be a good indicator to determine if a user is adopted or not.



2.)**Email domain(plot top right)** – Based on analysis we identify that the majority of users belong to below user domains and the remaining are categorised as others.

'gmail.com','yahoo.com','jourrapide.com','cuvox.com','gustr.com','hotmail.com'

3.) Creation Month- (plot bottom left) After analysis of the month of account creation we do see a good difference in the proportion of users who are Adopted and Non-Adopted, especially months 4,5 and 6 which is April,May and June



4.) Difference between Creation Time and Last Login Session Time (plot top right). This is an important feature we notice that people who have this value to be 0 (which means the they did not login after creating account) or 1 (which means that that after creating their account and they used the account only for a month) ,the adopted Users keep increasing when their engagement time is longer as shown in the graph below

Based on the plots alone we can identify important factors for prediction to be

1.) Difference between Creation Time and Last Login Session Time, more months of engagement with Asana more likely user is to Adopt
2.) If the user created the account in month of April/May(more likely for Non-Adoption) vs June-Oct(more likely for Adopton)
3.) If creation source, is personal_projects, the user is not likely to Adopt whereas if the creation source is via guest_invite or google_signup, they are more likey to adopt than other creation sources.
4.) If the email domain is Hotmail or gmail, the user is more likely to adopt than other user domains

Assumption:

1.) We are not sure that the engagement information can be used for prediction of Adopted Users. So we use two models one with engagement information and one without it (more details below)

Statistical Analysis:

After some statistical modelling we use indicators such as p-value to identify most important parameters, We use those parameters to run a Logistic Regression model. We get a higher accuracy when we use engagement information (which is difference between creation time and last login time) of 97% but when we remove it we get an accuracy of prediction of around 61%

But using the model we identify the most important features to be

1.) ‘creation source’ with all 5 categories included

2.) and if the account was created in month 4 (April) and month 03 which is March from the ‘creation_time’ variable

3.) If the user is 'opted_in_mailing_list' is also an important predictor

4.) If we are able to use engagement information, we can use the difference in time between account creation and last session creation column and use that in our model for more accuracy for prediction of Adopted Users or not.