Relax Challenge

First step:

Checking the datasets:

The users table conatins data of 12000 users and how their account was created (source) wheras the engagement table has 207917 entries about how the users visited the site. I quickly checked that the ‘visited’ column has no entry other than 1. I also checked the two datasets for discrepancy:

There are 8823 users who have ever used the application.

There are 12000 signed up for the application.

There is a discrepancy of 3177 users.

Second step:

Defining an "adopted user" as a user who has logged into the product on three separate days in at least one seven­day period,

I did two calculations:

1. We have analyzed logs for 8823 unique users. 18.77% of users who have used the application can be considered 'adopted users'.
2. 13.80% of all users who have signed up for the application can be considered 'adopted users'.

Third step:

I checked how the creation sources are divided:

creation\_source feature:

ORG\_INVITE 4254

GUEST\_INVITE 2163

PERSONAL\_PROJECTS 2111

SIGNUP 2087

SIGNUP\_GOOGLE\_AUTH 1385

opted\_in\_to\_mailing\_list feature:

0 9006

1 2994

enabled\_for\_marketing\_drip feature:

0 10208

1 1792

Fourth step:

At this point I started preparing for modelling and figuring out features which will help in modelling. The goal is to determine factors which will predict user adoption.

1. I dropped columns : last\_session\_creation\_time, invited\_by\_user\_id as these two columns have too many null entries
2. I dropped columns which are more descriptive in nature than predictive like, name, email, creation\_time etc.
3. I used pd.get\_dummies to fill creation\_source column with dummies ,if needed.

Fifth step:

In this step I applied Random Forest classifier with Gridsearch.

1. I did a train-test split
2. Since there is imbalance in the adoption field, I used the class\_weight = ‘balanced’ parameter of the classifier.
3. I got a training score of 0.84 and testing score of 0.72
4. With GridSearch, I could get upto 0.75 accuracy score.
5. Here is the orcer of feature importance-
6. feature\_importance

org\_id 0.935802

creation\_source\_PERSONAL\_PROJECTS 0.015795

enabled\_for\_marketing\_drip 0.015736

opted\_in\_to\_mailing\_list 0.014804

creation\_source\_SIGNUP\_GOOGLE\_AUTH 0.006451

creation\_source\_GUEST\_INVITE 0.005424

creation\_source\_SIGNUP 0.003120

creation\_source\_ORG\_INVITE0.002869