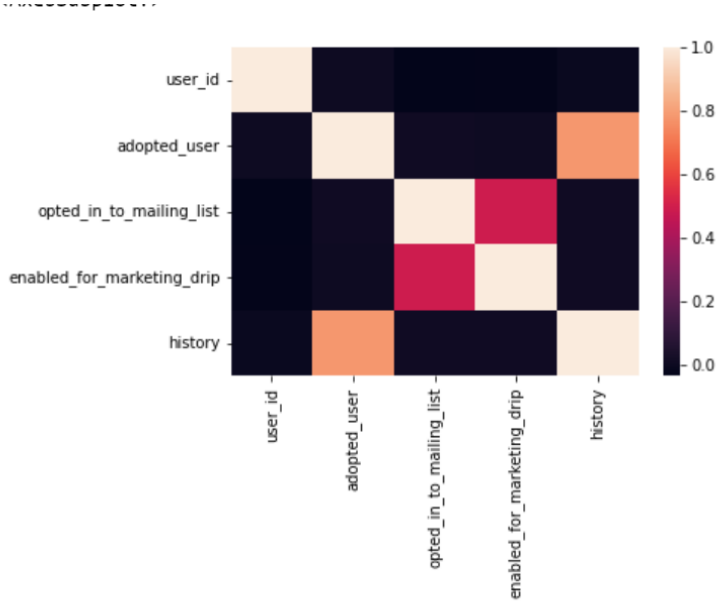
Relax Data Science Write Up

By: Wasinee Siewsrichol

I first started by cleaning up the dataset from the takehome_user_engagement.csv which had all of the user_id with the time they logged in. I then defined an "adopted user" as a user who logged into the product on 3 separate days in at least a 7 day period by using a function with timedelta from datetime. There were 1656 adopted users out of 8823 users.

Then I moved onto cleaning the second dataframe from takehome_users.csv which had 12,000 users who signed up within the past 2 years. There were some null values from last_session_creation_time and invited_by_user_id. I decided to drop the latter column since it had too many null values, and kept the first one to do further investigation. I also dropped name, email, and org_id which I deemed irrelevant to the target which is finding out which feature most affects if the user is adopted or not. I created a column named history which is the number of days the account has been accessed since its creation and then dropped the columns: creation_time' and 'last_session_creation_time'. I created a dummy variable for the categorical column and created a correlation heatmap using seaborn.



As you can see above, there is some sort of correlation between adopted user and history. More investigation needed to be done and so I did a train test split with a test size of 20%. I fitted the training datasets to the Random Forest Classifier and predicted the adopted user. The accuracy test and classification report is below.

```
Accuracy of test set was 0.9580736543909348
              precision    recall  f1-score   support

           0       0.97      0.98      0.97      1417
           1       0.92      0.87      0.89       348

    accuracy                           0.96      1765
   macro avg       0.94      0.92      0.93      1765
weighted avg       0.96      0.96      0.96      1765
```

As you can see, the Random Forest had a pretty high accuracy score of 96% with most scores hovering over the 90%. I also wanted to see the features which are most important below.

	feature	importance
3	history	0.886496
0	user_id	0.093253
1	opted_in_to_mailing_list	0.004379
2	enabled_for_marketing_drip	0.003403
4	creation_source_GUEST_INVITE	0.002983
5	creation_source_ORG_INVITE	0.002776
7	creation_source_SIGNUP	0.002408
8	creation_source_SIGNUP_GOOGLE_AUTH	0.002241
6	creation_source_PERSONAL_PROJECTS	0.002061

The highest feature importance is history, followed by user_id and the rest are pretty unimportant. These results do not surprise me but it is nice to see that the Random Forest did a pretty good job in predicting adopted users. For future improvements, I would like to do some hyperparameter tuning and use other types of models to fit the data to see which model works best. SInce I was trying to fit everything in about a 2 hour time frame as recommended by the prompt.

Github Jupyter Notebook Link:
https://github.com/WasineeSi/Springboard/blob/a349621e6da08463b9cbbb6fd1dcec25a75b0c9b/relax_challenge/relax%20challenge.ipynb