Relax Take Home Report:

What factors predict future user adoption?

The factors from the dataset that best predict user adoption are: ***last_session_creation_time, enabled_for_marketing_drip, creation_source - guest invite, org_id, creation_source - sign up.***

How did I get there?

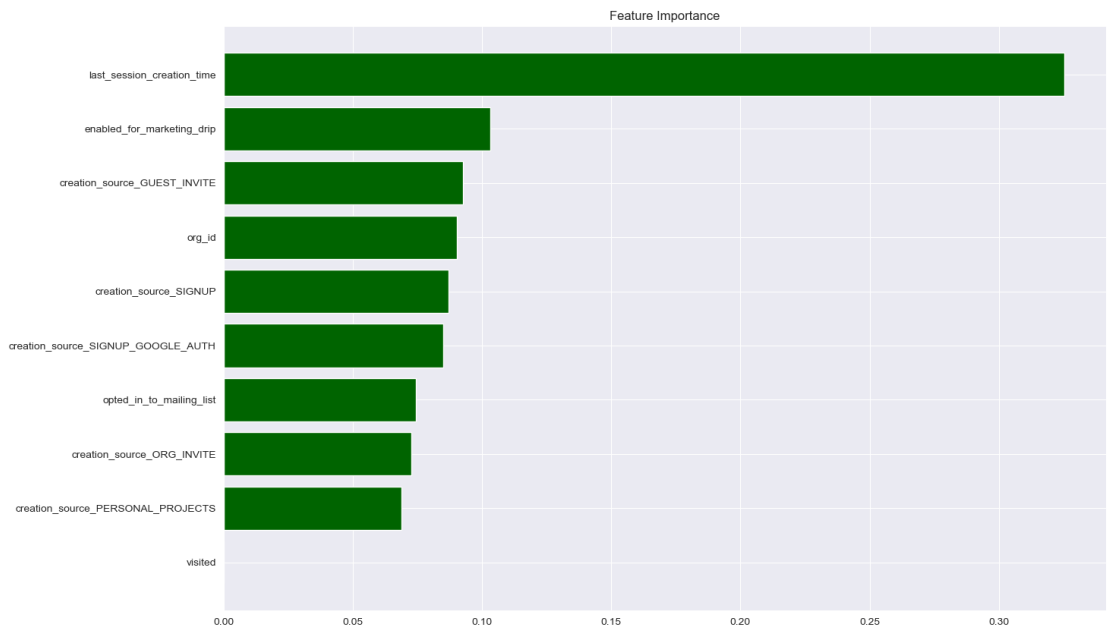
After merging both datasets together on matching user IDs, I defined an adopted user using the function below and added the results to the dataset:

```
merged_df.sort_values(by=['user_id', 'time_stamp'], inplace=True)
merged_df['adopted_user'] = 0

def check_adopted_user(group):
  group.sort_values(by='time_stamp', inplace=True)
  time_diff = group['time_stamp'].diff()

  if any(time_diff.dt.days <= 7):
    group['logins_in_seven_days'] = time_diff.rolling(window=7).count()
    if (group['logins_in_seven_days'] >= 3).any():
        group['adopted_user'] = 1
    else:
        group['adopted_user'] = 0
  else:
    group['adopted_user'] = 0
return group

merged_df = merged_df.groupby('user_id').apply(check_adopted_user)
```

Once adopted users had been defined, I explored class distribution (results were 18.5% adopted, 81.5% not adopted), and prepared the data for modeling via an XGBoost model. The goal was to model then plot feature importance to best understand which data features predict user adoption status.

I dropped the users name, email and account creation time, and encoded the creation source feature for use in modeling. Why did I drop creation time? After splitting creation time into separate features for day, month and year, every model I used was practically only using those three features and nothing else. I knew something was off and decided to proceed without them.

Here are the results of modeling and plotting feature importance:



Shown here are the results I stated above. The XGBoost model I used determines feature importance by considering the contributions of each of these features to the model's performance. It assesses how each feature contributes to the model's accuracy and how often each feature is used by the model across all boosting rounds. This indicates that the last session creation time contributed the most to predicting user adoption, followed fairly equally by being enabled for marketing drip, being invited to an organization as a guest, and the group of users they belong to.