Link to Github that contains the Jupyter Notebook:
https://github.com/SamayeLohan/Asana-Data-Science-Challenge

In order to figure out which features are likely indicators of future adoption or simply, which factors predict user adoption, I took a three-step approach that consisted of extracting the top values from a correlation matrix, applied univariate selection to the statistical test, and finally trained the feature matrix using a random forest classifier in order to yield the correlation factor for those features.

Preprocessing and EDA:

1) The initial step was to aggregate the data given using the two tables and define another column that represented the adopted user: Type: Boolean
    a) Adopted_user = 1 -> logged into the product on three separate days in at least one seven-day period
2) Combine the two data frames on *user_id* and calculate the following metrics:
    a) how much time it takes the user from creating an account to logging in for the first time
    b) Classify the email category into 5 groups (largest)
    c) Classify the organization category into small, medium, and large

These steps were used to construct a correlation matrix, shown below, which illustrated that the *last_session_creation_time* and the *diff_user_login_avg* have the highest correlation with respect to the adopted_user.

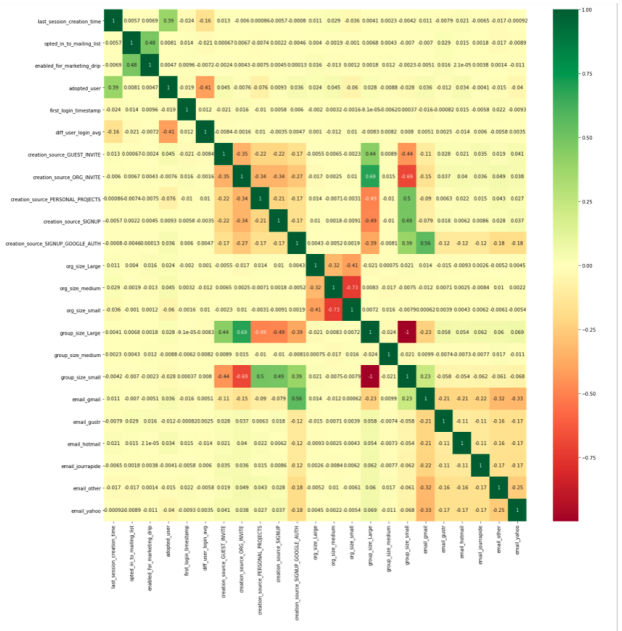

Figure 1: Correlation Matrix for the Adopted User

Univariate Selection - KBest Features:

Next, I performed a simple univariate calculation of the feature variables to visualize which ones had the strongest correlation. I used the SelectKBest module part of SciKit-Learn and chose the *chi2* score_func in order to determine the correlation:

We can see that the top 4 features in determining an adopted user are *diff_user_login_avg*, *creation_source_PERSONAL_PROJECTS*, *last_session_creation_time*, and *org_size_small*.

Output 10 best features using SelectKBest

```
In [60]: print(featureScores.nlargest(10,'Score'))
```

	Specs	Score
4	diff_user_login_avg	2716.483567
7	creation_source_PERSONAL_PROJECTS	56.844560
0	last_session_creation_time	36.239314
12	org_size_small	22.581597
5	creation_source_GUEST_INVITE	20.283149
21	email_yahoo	15.237361
11	org_size_medium	15.133797
9	creation_source_SIGNUP_GOOGLE_AUTH	13.848040
18	email_hotmail	12.762762
16	email_gmail	11.048830

Figure 2: Univariate Selection using SelectKBest

Training the Data:

To solidify our understanding of the important features we have gathered so far, I took the problem one step deever by using a *RandomForestClassifier* to further examine the feature matrix. The reason for choosing a tree-based classifier is because they are very powerful when dealing with a large dataset that has a lot of variance and independence. I used an 80/20 split in order to give importance to the training set:

I was able to achieve an AUC Score of .98 which is really good because we want our curve to be as close to the upper left corner as possible and achieving a score of 0.98 means that our model performs really well. Figures 3 and 4 extend our previous knowledge on the most important features in predicting user adoption. Figure 4 shows the *feature_importances_* scores for the different feature variables. In conclusion, the top features are: *diff_user_login_avg*, *last_session_creation_time*, *first_login_timestamp*, and *opted_in_to_mailing_list*.

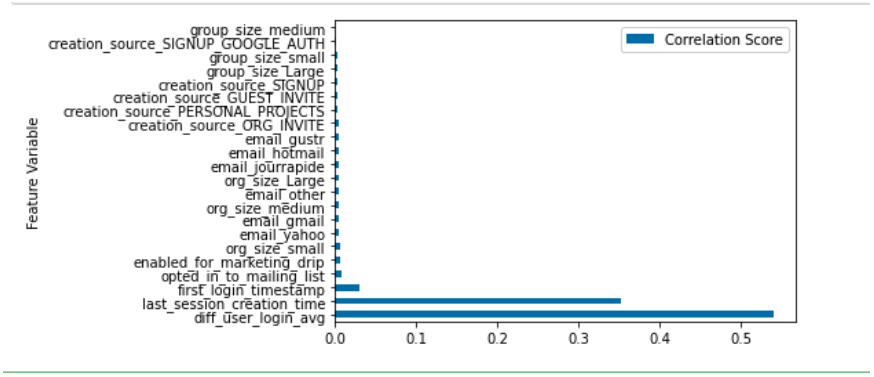

Figure 3: Bar Plot of Feature Importance Scores

```
diff_user_login_avg                 0.540887
last_session_creation_time          0.351690
first_login_timestamp               0.030292
opted_in_to_mailing_list            0.008182
enabled_for_marketing_drip          0.006819
org_size_small                      0.005713
email_yahoo                         0.004770
email_gmail                         0.004747
org_size_medium                     0.004720
email_other                         0.004712
org_size_Large                      0.004371
email_jourrapide                    0.004208
email_hotmail                       0.004207
email_gustr                         0.003984
creation_source_ORG_INVITE          0.003949
creation_source_PERSONAL_PROJECTS   0.003666
creation_source_GUEST_INVITE        0.003656
creation_source_SIGNUP              0.002853
group_size_Large                    0.002559
group_size_small                    0.002044
creation_source_SIGNUP_GOOGLE_AUTH  0.001969
group_size_medium                   0.000003
dtype: float64
```

Figure 4: Feature Importance Scores