Relax Challenge

I discovered that the last_session_creation_time feature and the invited_by_user_id feature had some missing data. The invited user column missing values were not filled, but rather replaced with a new feature based on if they were invited or not. The last session time was set equal to the initial creation time for those missing values. No other issues were noted with the data.

```
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   object_id                   12000 non-null  int64
 1   creation_time               12000 non-null  object
 2   name                        12000 non-null  object
 3   email                       12000 non-null  object
 4   creation_source             12000 non-null  object
 5   last_session_creation_time  8823 non-null   float64
 6   opted_in_to_mailing_list    12000 non-null  int64
 7   enabled_for_marketing_drip  12000 non-null  int64
 8   org_id                      12000 non-null  int64
 9   invited_by_user_id          6417 non-null   float64
```

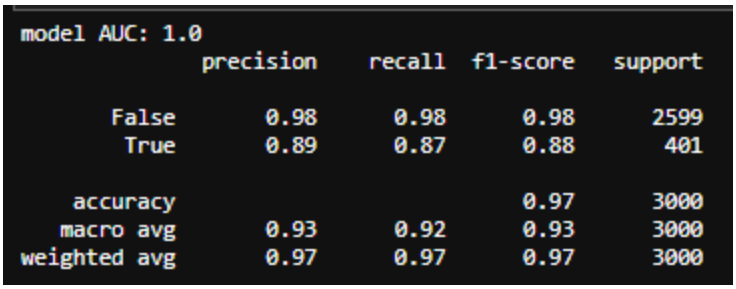
Figure 1: Missing data by column

For feature engineering, the company name was stripped from the email address and put into 1 of 7 categories (top 6 companies or other), and to reduce the number of dates/ timestamps a time between the last session and the creation date was created while also truncating the creation date to only a month and year. The email address, name, last session date, and organization id features were all deleted because they were unique identifiers. The object id was also eliminated after it was used to join the predictor feature table to the adopted user data table.

```
Top 10 company email addresses
gmail.com         3562
yahoo.com         2447
jourrapide.com    1259
cuvox.de          1202
gustr.com         1179
hotmail.com       1165
rerwl.com            2
oqpze.com            2
qgjbc.com            2
dqwln.com            2
Name: email_company, dtype: int64
```

Figure 2: Top 6 email address companies

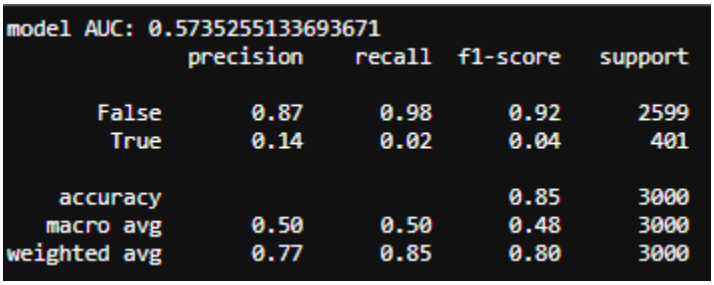
To prepare for the model, the categorical features were one-hot encoded, and then the data was split into training and test data such that 25% of the data was reserved for testing. A random forest model was chosen with default hyperparameters since it is typically a good out-of-the-bag predictor model. The model scores were an AUC of 1.0 and model accuracy of 0.97. The ranked list of feature importance was essentially only a single feature, with the time between the last session and the profile creation date having a weight of 0.886. I found this highly suspicious and realized that although the feature is not technically confounded with the target variable, the longer someone has had their account increases the likelihood that there was a week that they used the software three times. This information also is not particularly helpful for the company. A model was then ran without that feature, but the model accuracy was worse than if we predicted that everyone was not an adopted user.



model AUC: 1.0

	precision	recall	f1-score	support
False	0.98	0.98	0.98	2599
True	0.89	0.87	0.88	401
accuracy			0.97	3000
macro avg	0.93	0.92	0.93	3000
weighted avg	0.97	0.97	0.97	3000

Figure 3: Model performance with time between last session and creation



model AUC: 0.5735255133693671

	precision	recall	f1-score	support
False	0.87	0.98	0.92	2599
True	0.14	0.02	0.04	401
accuracy			0.85	3000
macro avg	0.50	0.50	0.48	3000
weighted avg	0.77	0.85	0.80	3000

Figure 4: Model performance without time between last session and creation