relax_report

March 3, 2018

1 Relax Inc. Analysis

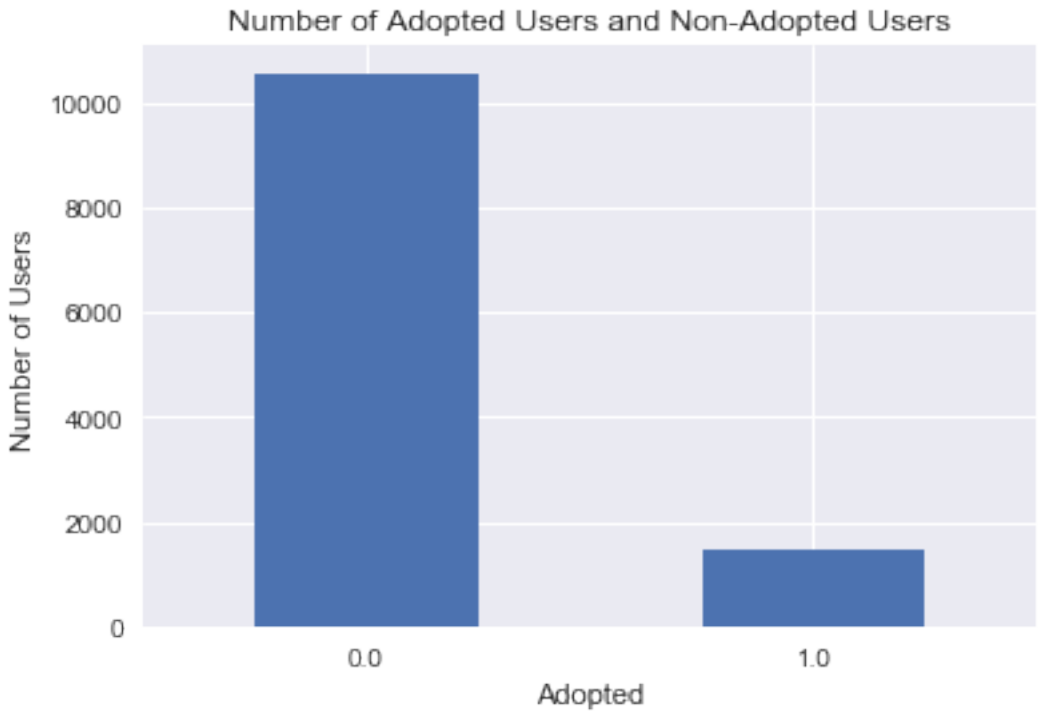
Problem: Determine the factors that predict future user adoption.

1.1 User Adoption

User adoption is described as "a user who has logged into the product on three separate days in at least one seven-day period." This was calculated by creating a pivot table of logins and users and then resampling the data by 1 week intervals. If the user had more than 3 logins in one of the weeks, then the user was considered adopted. There are less than 1500 adopted users in the dataset, which means there are ~1050 non-adopted users, as seen in the plot below.

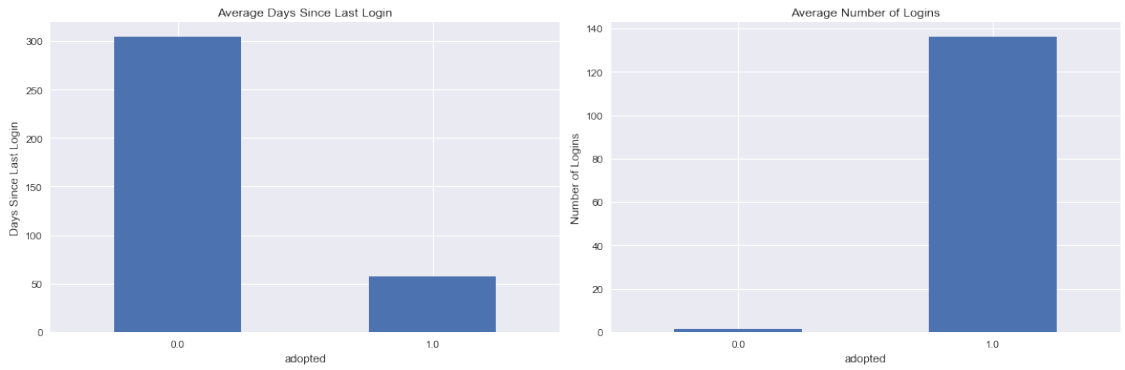
1.2 Attributes Added

The Relax Inc. dataset contained a last session created, which was used to create a feature called days since last session. By taking the newest date in the column and subtracting the value, we can get the days since the last session. The next feature created was the domain of the email address provided. This was extracted from each user in the dataset. The total number of logins from a user was also calculated for each user. The last two attributes created were based on whether the user was invited by another user and if the user was part of an organization.



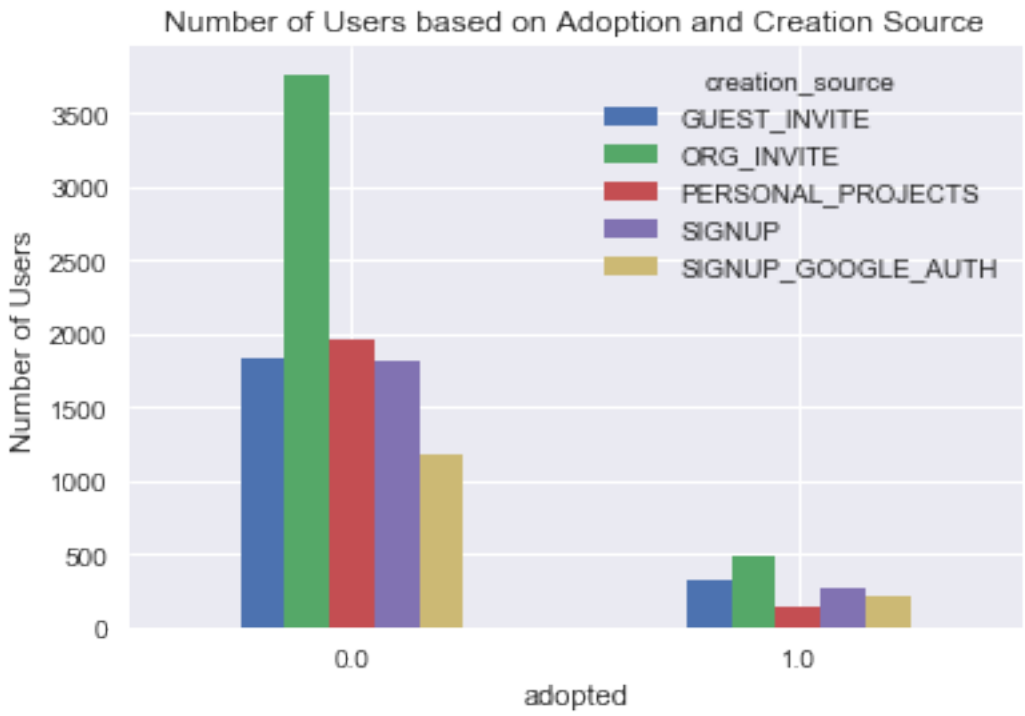
1.3 Important Features

When you break down adopted users and non-adopted users by the average number of days since the last login, there is a big difference between the two (300 for non-adopted vs ~50 for adopted). The same goes for the average number of logins for adopted (~140) vs. non-adopted users (< 10). These two features are the most important.

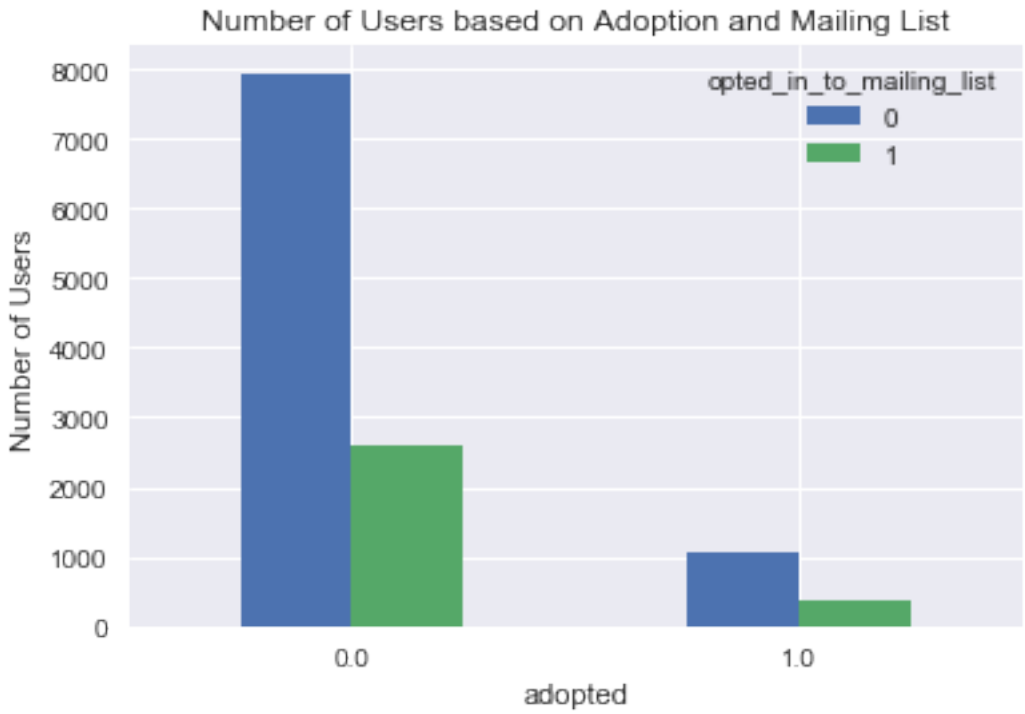


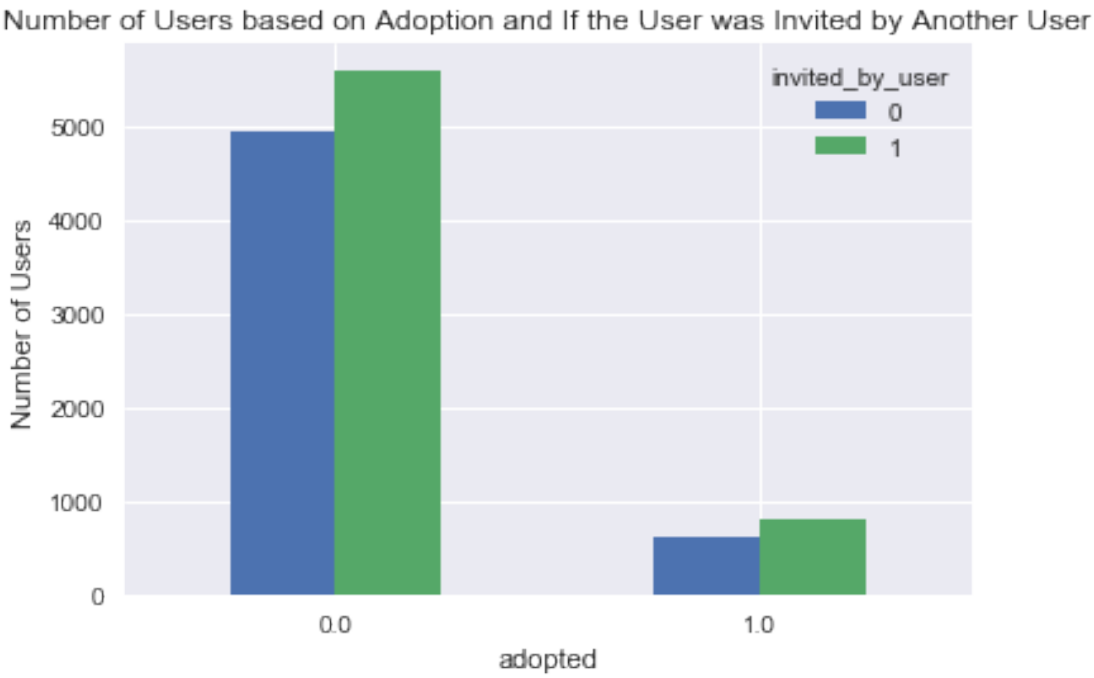
The Creation Source seems to have a small importance for adoption. Personal projects seem be the highest predictor of whether a user adopts the product because there is a stark difference between the adopted users and non-adopted users.

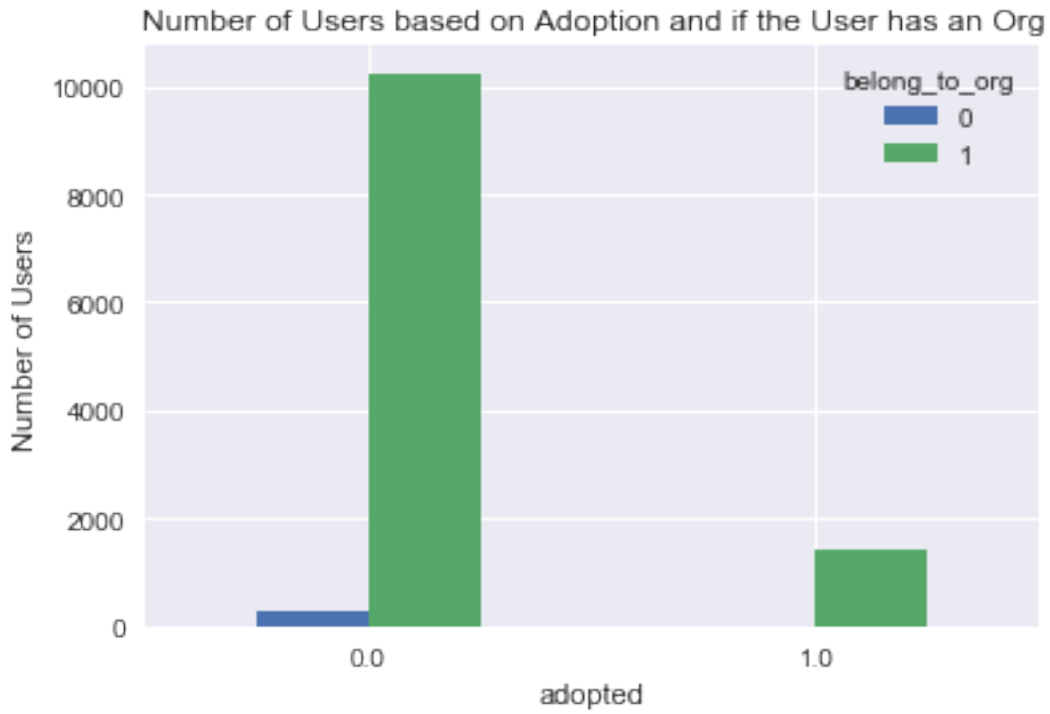
```
<matplotlib.figure.Figure at 0x10ef3bc18>
```



The next three plots based on whether a user was part of the mailing list, if the user was invited by another user, and if the user was part of an organization did not show as much importance like the number of logins or the days since the last login. There is a big difference in adoption status for both opting in and out of the mailing list. There is also a large difference in adoption status between users that were invited by other users and those that were not. The org plot does not actually show anything important.


Number of Users based on Adoption and Mailing List
opted_in_to_mailing_list
0
1
Number of Users
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
1.0
adopted


Number of Users based on Adoption and If the User was Invited by Another User
invited_by_user
0
1
Number of Users
5000
4000
3000
2000
1000
0
0.0
1.0
adopted



It was difficult to view the email domains in one plot as there were over 1000 email domains. The output of the table would be too large to view all of the columns as well. The same goes for the organization ids. These were not included in this report but they could ultimately be important features.

A Random Forest model was created to predict the whether a user will adopt the product or not. The model's accuracy was 97%, with a precision of .97. The recall was also .97 and the ROC AUC was .87. The model does really well in predicting whether a user will adopt the product or not. Random Forests have a nice feature that shows how important they are. The top 10 important features are shown below.

	features	importance
2	logins	0.523156
3	days_since_last_login	0.205436
6	creation_source_PERSONAL_PROJECTS	0.015259
1536	email_domain_@yahoo.com	0.007591
5	creation_source_ORG_INVITE	0.007156
8	creation_source_SIGNUP_GOOGLE_AUTH	0.006977
717	email_domain_@gmail.com	0.006753
4	invited_by_user	0.006520
7	creation_source_SIGNUP	0.005967
0	opted_in_to_mailing_list	0.005903