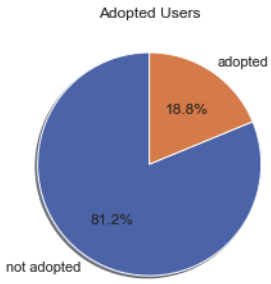
Relax Challenge

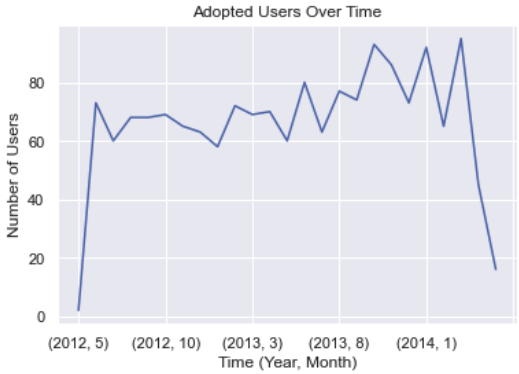
Problem: Identify which factors predict future user adoption.

My approach to answer this problem was to first create a feature to track user adoption. To do so I used the takehome_user_engagement.csv and pandas to manipulate the data. This data was then joined with the takehome_users.csv. I also created month and year features for when the account was created and a feature for the type of email that was used. The time features created could be especially useful if there are abnormal spikes in the number of adopted users.



After cleaning the data, I wanted to visualize the distribution of the adopted users and the number of adopted users over time. The total number of adopted users is only a small portion of the total users of the app. There does not appear to be any seasonal trends in the number of adopted users over time, but there are some months with larger jumps and drops and it would be interesting to see if there were significant events around these times such as updates to the app.

When looking for variables, I used various Chi-square tests of independence to compare different categorical variables with adopted users. The null hypothesis of this test is that these two variables occur independently of each other. These are the variables that I tested adopted users against:



- Creation_source
- Created_year
- Created_month
- Opted_in_to_mailing_list
- Enabled_for_marketing_drip
- email_provider

Each of these tests resulted in a p-value of ~0.000 except for opted_in_to_mailing_list and enabled_for_marketing_drip which had a p-value of ~0.502 and ~0.715 respectively. This results in a failure to reject the null hypothesis for opted_in_to_mailing_list and enabled_for_marketing_drip meaning that these variables are likely independent of one another. For the other variables, it results in a rejection of the null hypothesis meaning that these variables are likely correlated with adopted users.

It would be interesting to look more into the dates of when the users became adopted users and this is something I thought would be useful after the fact. I feel that these time trends would be a better feature to look into rather than the year created or month created as it takes time before a user becomes an adopted user. I also think that tree based models will perform better during the modeling phase, due to the fact that most of the variables are categorical. It would also be interesting to view the feature importance generated during modeling.