To create a predictive model for adopted users, first I had to create a column for whether the user was “adopted.” In order to do this, I used a series of SQL queries, first to narrow the user base down to only the users who had visited at least three times at all (Query 1)

This narrowed the user base from 12,000 to only 2,248 that could plausibly be adopted users. I discovered that although these were only 19% of users, they represented 97% of logins. It also turned out that 26% of users who created an account never logged in at all.

I also discovered that although the average user logged in 17.3 times, if you include only users who logged in at least once, the average becomes 23.6 times. If you include only users who logged in at least three times, you reach an average of 89.4 logins per user.

After indexing, I ran two SQL queries to find the adopted users. First, I found all logins by the same user within one week with a self-join (Query 2), and then, with another self-join, I found all sets of three logins within one week (Query 3). This turned out to include 1602 users. I added the target feature to the users column as either 0 for non-adopted users and 1 for adopted users.

Before creating my models, I added two more predictive features: One that turned the account creation date into an integer (number of days after the first account creation), and one that counted the days between account creation and most recent login. Although there were too many values of organization ID to make categorical, and I didn’t think it made sense to treat them as integers (since the ordering was relatively arbitrary), I created a binary variable for whether they were in one of the most common organizations (15% were). I also tried to create a variable for whether they were invited, but that turned out to be redundant with the dummy variables I created for account source, so I disregarded it.

When I created a Random Forest model, the predictive value was astonishingly high, with every average scoring metric at least 0.95. When I looked at the graph of important features, it quickly became clear why. (Figure 1) It turned out that the number of active days was nearly the only significant predictive feature. This was unsatisfying to me, as I didn’t feel like it was knowable in advance, so I tried taking that feature out and creating a new model.

The second Random Forest model relied almost entirely on the date of account creation (Figure 2). This was a bit better, but still unsatisfying. However, when I tried taking that feature out too, I got a completely nonfunctional model that exclusively predicted “No.”

A K Nearest Neighbors model tuned to use just 1 neighbor did nominally better by predicting 26 positive results, but it should’ve predicted 326, so it still wasn’t very good.

From this, I concluded that the dates really were the only predictive feature, so I created a decision tree with hyperparameters chosen for interpretability (Figure 3), and called it a day.

Query 1:
'SELECT user_id,
SUM(visited) AS visits
FROM engagement
GROUP BY user_id
HAVING visits >= 3'

Query 2:
"'SELECT x.user_id,
x.time_stamp AS first_stamp,
y.time_stamp AS second_stamp
FROM engagement as x
JOIN engagement as y
ON x.user_id = y.user_id
WHERE x.user_id IN
(SELECT user_id FROM active)
AND x.time_stamp < y.time_stamp
AND y.time_stamp <= DATE(x.time_stamp, '7 days')"'

Query 3:
"'SELECT x.user_id,
x.first_stamp,
x.second_stamp,
y.second_stamp AS third_stamp
FROM twotimes AS x
JOIN twotimes AS y
ON x.user_id = y.user_id
WHERE x.second_stamp < y.second_stamp
AND y.second_stamp < DATE(x.first_stamp, '7 days')"'
Figure 1:

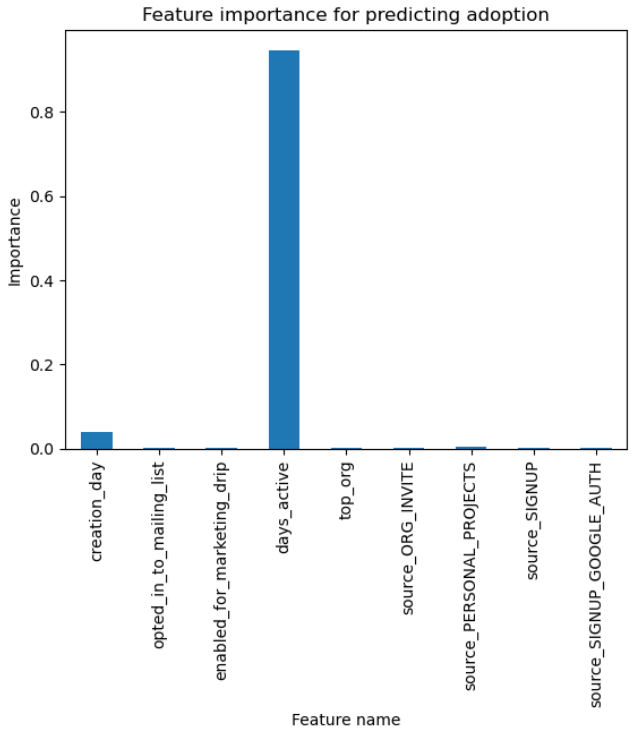

Feature importance for predicting adoption
Importance
0.0
0.2
0.4
0.6
0.8
creation_day
opted_in_to_mailing_list
enabled_for_marketing_drip
days_active
top_org
source_ORG_INVITE
source_PERSONAL_PROJECTS
source_SIGNUP
source_SIGNUP_GOOGLE_AUTH
Feature name

Figure 2:

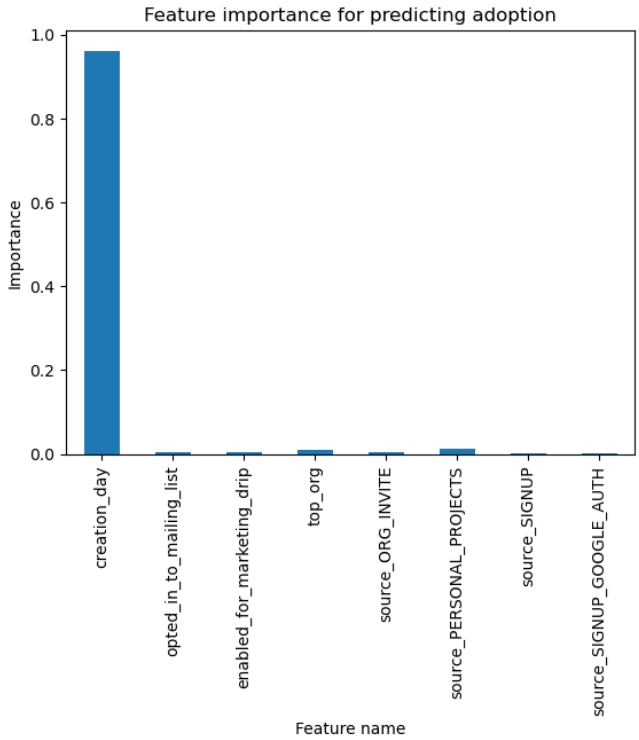
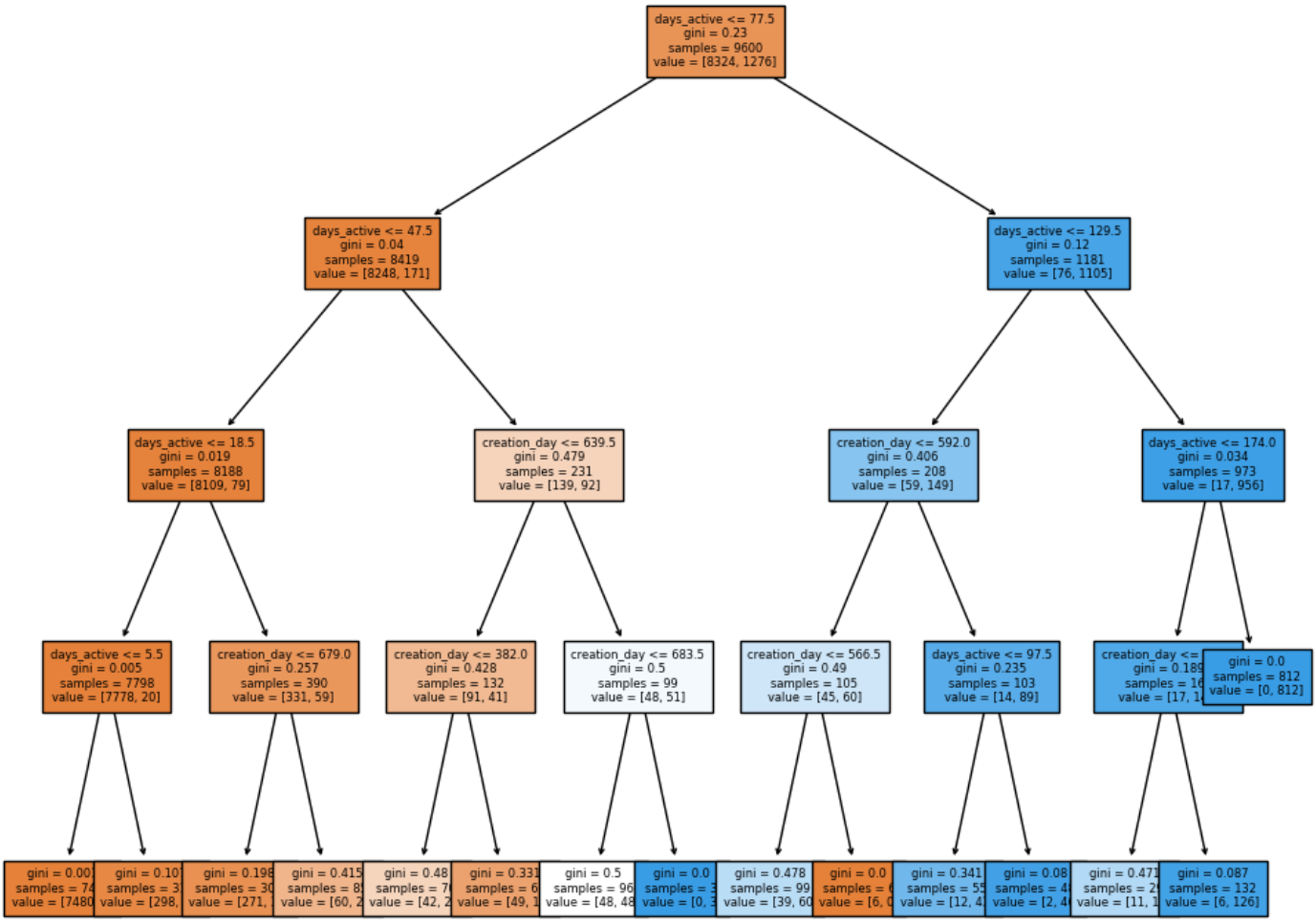

Feature importance for predicting adoption
Importance
0.0
0.2
0.4
0.6
0.8
1.0
creation_day
opted_in_to_mailing_list
enabled_for_marketing_drip
top_org
source_ORG_INVITE
source_PERSONAL_PROJECTS
source_SIGNUP
source_SIGNUP_GOOGLE_AUTH
Feature name

Figure 3:


days_active <= 77.5
gini = 0.23
samples = 9600
value = [8324, 1276]
days_active <= 47.5
gini = 0.04
samples = 8419
value = [8248, 171]
days_active <= 129.5
gini = 0.12
samples = 1181
value = [76, 1105]
days_active <= 18.5
gini = 0.019
samples = 8188
value = [8109, 79]
creation_day <= 639.5
gini = 0.479
samples = 231
value = [139, 92]
creation_day <= 592.0
gini = 0.406
samples = 208
value = [59, 149]
days_active <= 174.0
gini = 0.034
samples = 973
value = [17, 956]
days_active <= 5.5
gini = 0.005
samples = 7798
value = [7778, 20]
creation_day <= 679.0
gini = 0.257
samples = 390
value = [331, 59]
creation_day <= 382.0
gini = 0.428
samples = 132
value = [91, 41]
creation_day <= 683.5
gini = 0.5
samples = 99
value = [48, 51]
creation_day <= 566.5
gini = 0.49
samples = 105
value = [45, 60]
days_active <= 97.5
gini = 0.235
samples = 103
value = [14, 89]
gini = 0.0
samples = 812
value = [0, 812]