Hazen Relax Take Home Challenge

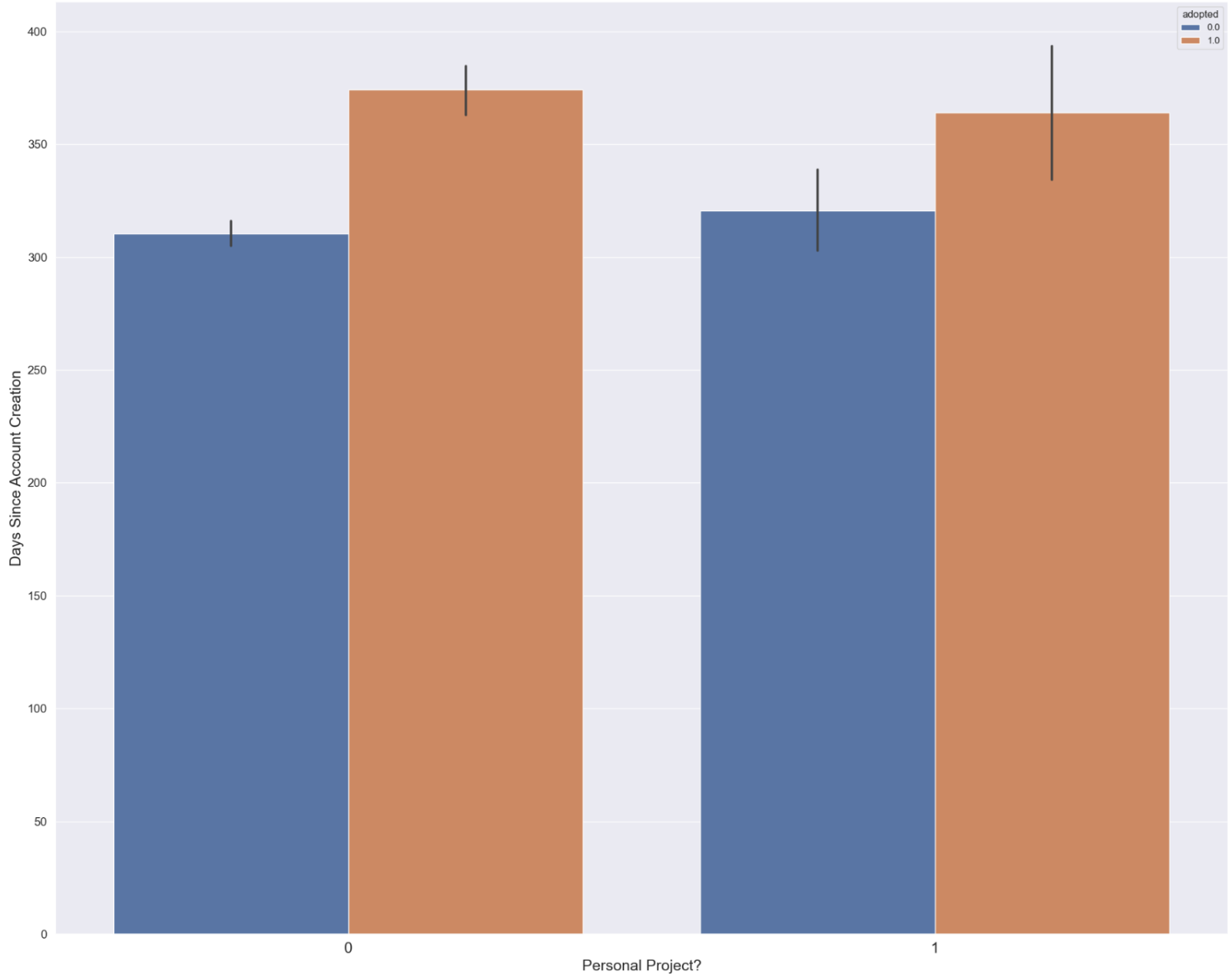
Datasets:
The data was encoded in "cp1252", a Windows encoding, so attention to reading the csv properly was made. The users dataset contained float64 columns, object columns, and int64 columns. The float64 columns needed to be recast as int64 in order to run them through any ML algorithms, and the object columns needed to be either dropped or one-hot-encoded to adjust.

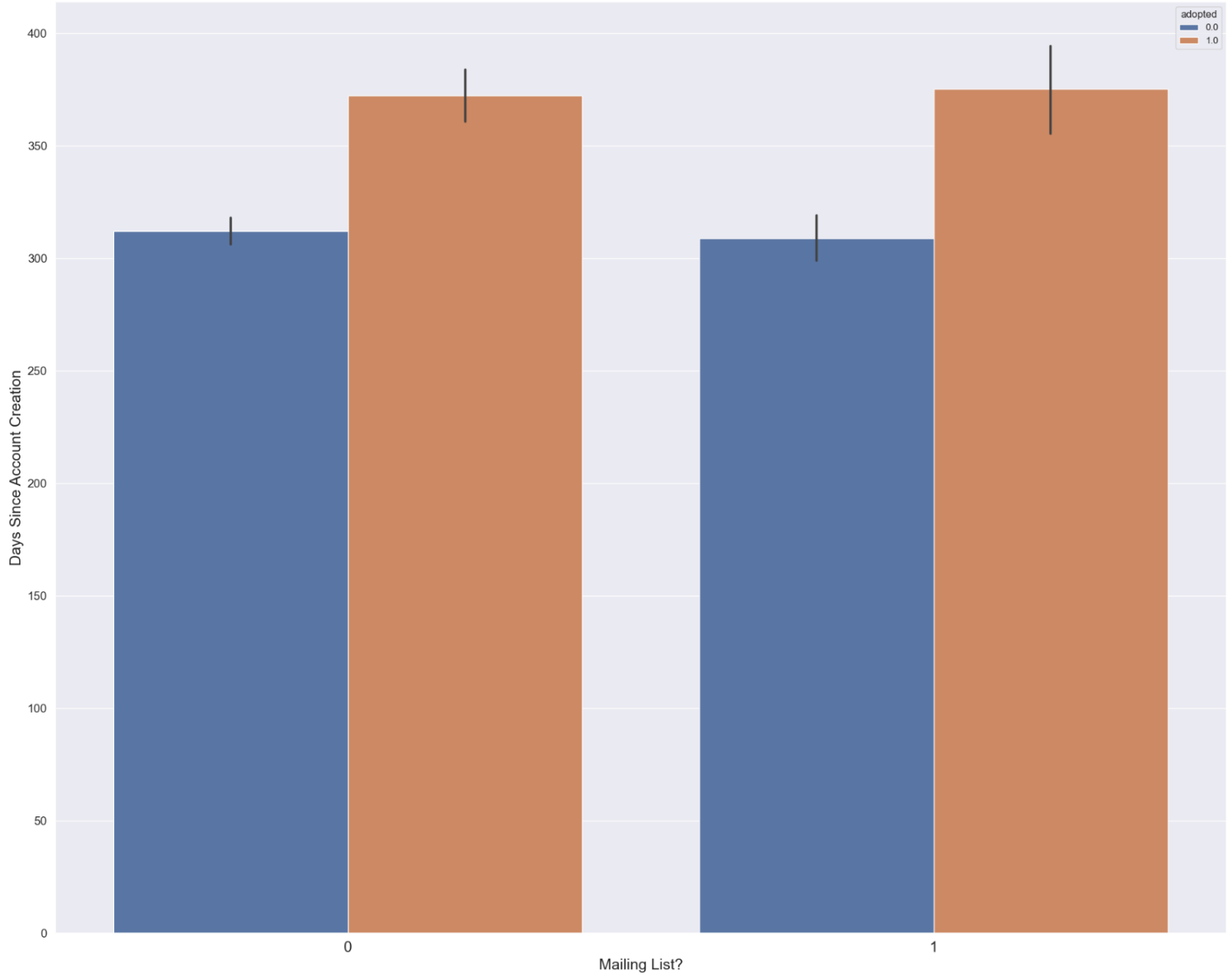
Data Wrangling:
The target variable was adopted users, and that column was calculated from grouping users, aggregating their logins for every 7 day window (a rolling count), and then counting every user who had at least one 7 day window with 3 or more logins. In the takehome_users data the object_id was user_id, so I had to switch over that column. The only two columns needed are the user_id and adopted binary classification, so I set the index to user_id (renamed as object_id) and dropped the other column in the rolling dataframe. I subsequently joined the new dataframe to the adjusted users dataframe with a left join. It was insightful to know how long users have been taking part, and accordingly I created a “number of days” column before dropping the last creation time column.

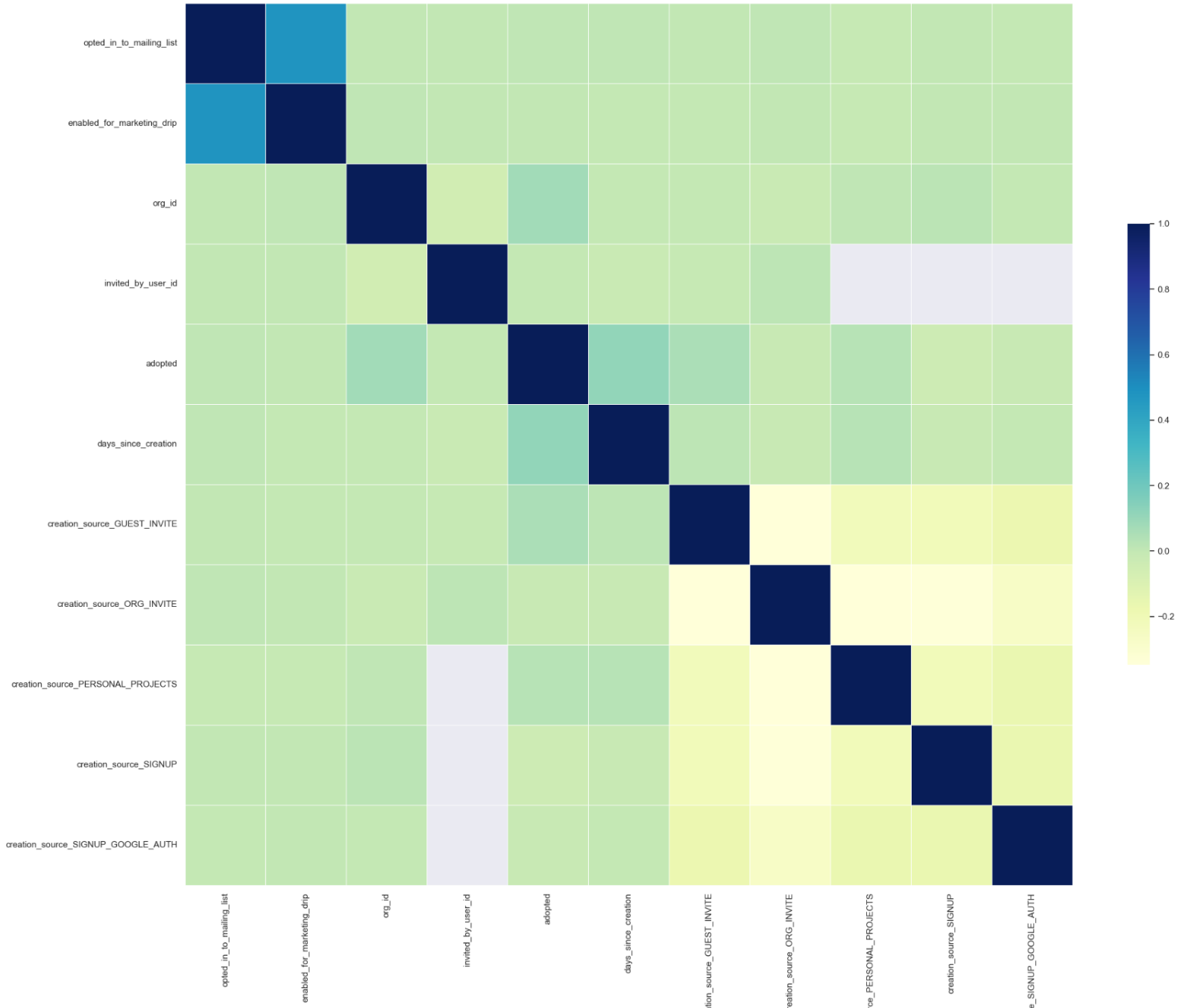
EDA:
Some correlations were present (see heatmap on the following page) but nothing strongly dominated the factors for adopted users. The adopted users had more days since creation of their accounts for both those on and not on the mailing list. If the creation source was a personal project, that did have an influence on both the number of days since creation and on adopted users. In the cross section between marketing drip and mailing list adoption, the effect on adopted users appeared to be minimal.

Predictions:
As this is a binary variable of adopted users, logistic regression was used to assess the accuracy and error. While the error was low, the accuracy was decently high at 86.5%. With recursive feature elimination, the following factors were top ranked in predicting adopted users with 5 features selected at a time.

- opted_in_to_mailing_list, Selected True, Rank: 1.000
- creation_source_GUEST_INVITE, Selected True, Rank: 1.000
- creation_source_PERSONAL_PROJECTS, Selected True, Rank: 1.000
- creation_source_SIGNUP, Selected True, Rank: 1.000
- creation_source_SIGNUP_GOOGLE_AUTH, Selected True, Rank: 1.000


adopted
0.0
1.0
400
350
300
250
200
150
100
50
0
Days Since Account Creation
0
1
Personal Project?


adopted
0.0
1.0
400
350
300
250
200
150
100
50
0
Days Since Account Creation
0
1
Mailing List?


opted_in_to_mailing_list
enabled_for_marketing_drip
org_id
invited_by_user_id
adopted
days_since_creation
creation_source_GUEST_INVITE
creation_source_ORG_INVITE
creation_source_PERSONAL_PROJECTS
creation_source_SIGNUP
creation_source_SIGNUP_GOOGLE_AUTH
1.0
0.8
0.6
0.4
0.2
0.0
−0.2