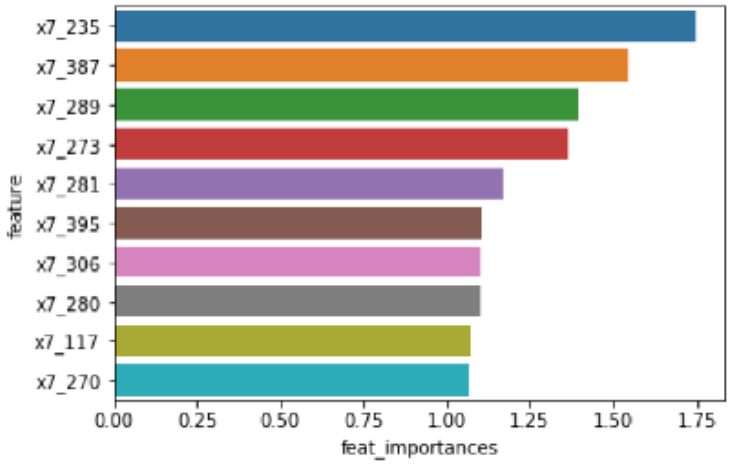
James Buckel

Relax Challenge Findings

Overall, it appears that the organization to which members belong is the most important feature to predict adopted users per our model. Organizations 235, 387, 289, 273, and 281 are the features with the strongest predictive power for our model, in that order.



Data Science Method Steps

I. Data Wrangling

After getting a sense of the data, I labeled the target feature and grouped and counted user visit time using the User Engagement Summary data. This was done by classifying users as using at least three days or more (1), or using for fewer days. I then merged the tables.

II. EDA

Next, I cleaned the data, dropping a few columns like User Names. I then filled missing values with the medians. From statistics and visualizations, I determined that there was significant class imbalance (adopted users are 17% of records). Furthermore, I realized that org_id will have over 400 variables after encoding which is less than ideal to handle.

III. Preprocessing/Modeling

I separated the target feature from the data. After scaling features and using one hot encoding for categorical variables, data was split into train and test sets. I tried logistic regression with mixed results. F1 score and recall were lower while precision was higher.

For future directions, the model may have been improved with more thorough feature selection and engineering, by balancing classes particularly for multiclass variables, and perhaps different model choice like RandomForestClassifier.