Relax Data Science Challenge - Report

After an initial review and exploration of the data provided, I ultimately found that a user's duration, defined as the number of days between account creation and last login, was the most influential feature in predicting user adoption. I will outline how I got to this conclusion and other considerations below.

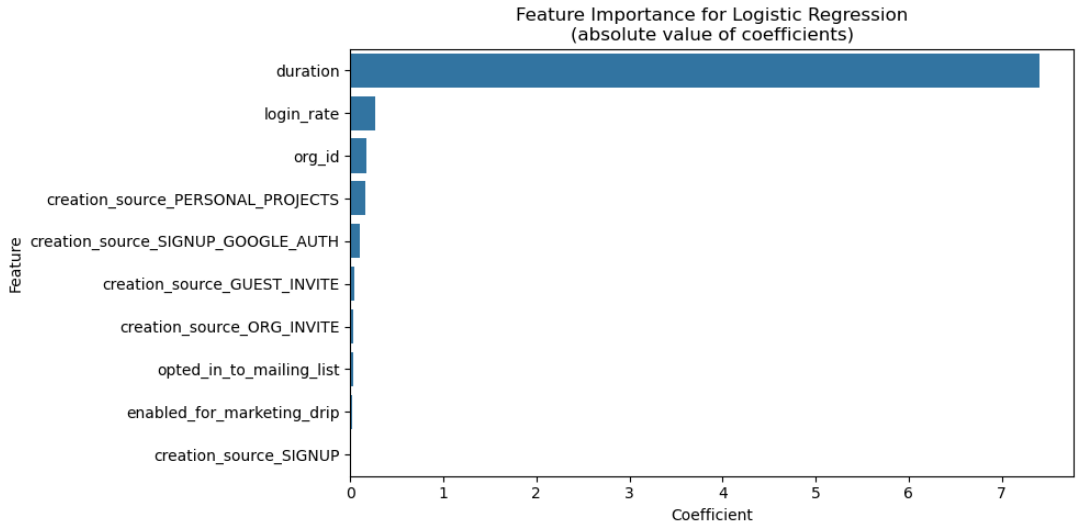
Using the user engagement data, I was able to define "adopted" users by ID numbers, which revealed 1,602 of the 12,000 total users had adopted the service. Of the features provided, IDs and most categorical variables (name, email) do not intuitively have influence on whether someone becomes an adopted user. However, it was interesting to see patterns in the creation source, such that users who signed up from an ORG_INVITE had the most total adoptees; however, on a percentage basis, GUEST_INVITE and SIGNUP_GOOGLE_AUTH. The creation source could be an important factor, and therefore I will create a dummy variable for each of these sources to determine their influence on predicting adopted users.

I also defined two additional variables to capture time and frequency of logins:

- duration = days between creation and last login
- login_rate = average number of logins per day (total logins per user / duration)

In the modeling stage, I wanted to try various algorithms that perform well with binary classifications, including logistic regression, K-nearest neighbors, support vector machine, decision tree, and random forest. Each model was measured on a 5-fold cross-validation F1 score and checked for overfitting by comparing F1 and accuracy scores run on train versus test data. Decision tree and random forest has the highest F1 scores, at around 94%. However, I worry these two models are overfitted because the evaluation scores on the training data are 100% accurate. Turning to logistic regression, KNN, and SVM models, these three resulted in about 89% cross-validated F1 scores. Unlike the tree algorithms, these models did not show the same overfitting patterns.

I selected the logistic regression model to evaluate the coefficients and feature importance, as shown in the figure below. This clearly shows duration by far has the highest influence on the predictions.



Future developments of this analysis would be to further refine the prediction model through hyperparameter tuning, as well as defining and engineering new features that could be influential.