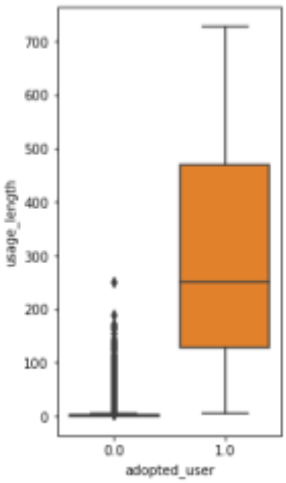
Relax Inc. Take Home Challenge
Project Report
Christopher Chung

GitHub

For this project, we used two datasets. The first dataset included data on all the users who signed up for the product and the second dataset included usage data with a row for each day that a user logged into the product. The objective was to identify which factors predict future user adoption, where user adoption is defined as a user who has logged into the product on three separate days in at least one seven-day period.

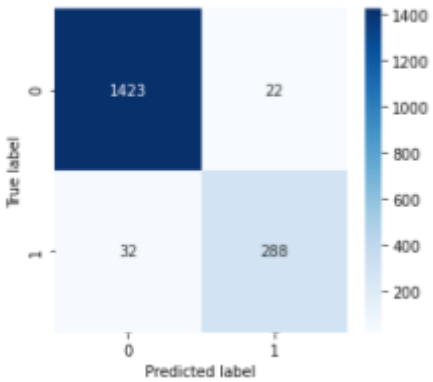
Figure 1: Usage length



We engineered a feature called 'usage_length', which represents the amount of time from account creation to the last recorded login. There was a significant difference in 'usage_length' between adopted users and non-adopted users, as shown in Figure 1, in that adopted users clearly had longer usage lengths.

We also examined the relationship between other features and our target variable, such as creation source, whether they opted into receiving marketing emails, whether they are on the regular marketing email drip, and whether the user was referred by another user. Most of these features showed no noticeable relationship, with similar behavior among adopted users and non-adopted users. These figures are shown on the next page.

Figure 2: Confusion matrix for logistic regression model

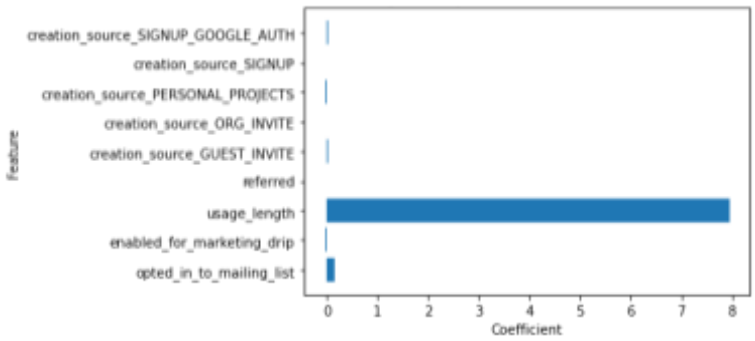


We trained and tested three classification models to predict user adoption. Our best performing model was the Logistic Regression algorithm, with a prediction accuracy of 97%. The confusion matrix for this model is shown in Figure 2.

In order to identify which factors predict future user adoption, we extracted the coefficients of each feature from our logistic regression model. The graph of the feature values is shown in Figure 3.

The most important predictor for user adoption, by far, was our 'usage_length' variable. This variable was engineered by subtracting the time of last recorded login by the account creation time, measured in days. The coefficient for this feature was positive and had a large magnitude of approximately 8. This means that an increase of 'usage_length' by one unit (scaled) increases the odds that the user will be adopted by 8 times.

Figure 3: Feature coefficients for logistic regression model



For future work, it would also be valuable to have demographic data on each of our customers as well as information on the organization that the user belongs to. This would be helpful to explore other possible customer segmentations.

Figure 4:

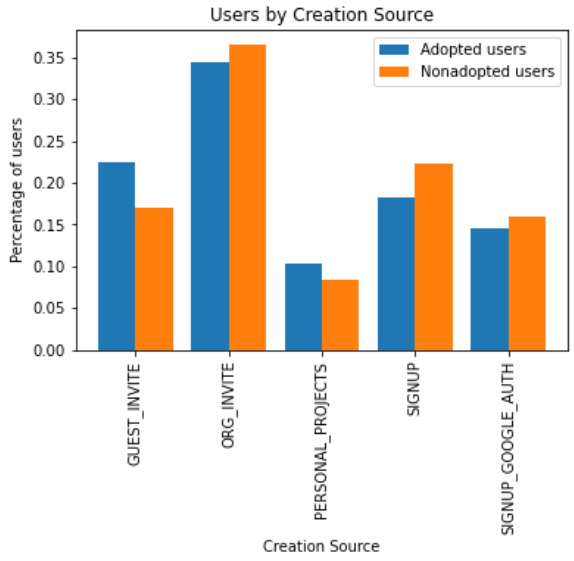


Figure 5:

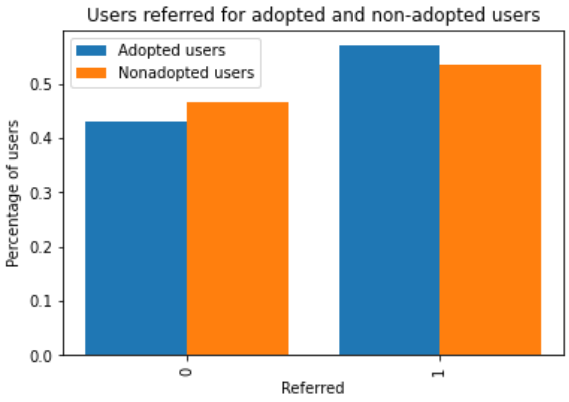


Figure 6:

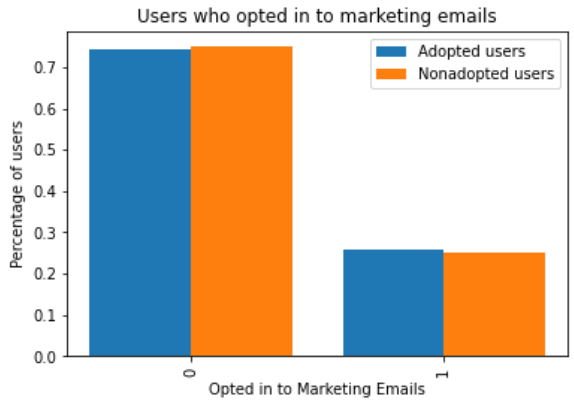


Figure 7: