Report for Relax Challenge by V. Moore

This exercise asks us to identify adopted users as those who log into a service on three or more days in any given week and then determine if any features associated with users enables us to predict who will become an adopted user.

The correlation heatmap provides a useful visual of relationships, but value of these interpretations is limited by the fact that correlation evaluation works better for continuous variables. Most comparisons of features with each other do not show strong relationships. Email domain shows very significant chi-squared relationships with creation source, invited status, and adopted status, as well as a fairly strong relationship with opting in to mailing list. Creation source shows very strong relationships with invited and adopted statuses and with days between last session and account creation dates. Opting in to the mailing list shows a very strong relationship with enabling for marketing drip. Organization ID shows a strong relationship with adopted status and interval between last session and account creation dates. Adopted status is strongly correlated with interval between last session and account creation dates, though this is not a surprise.

Next we will use machine learning to develop a model to predict adopted users. First, variables that are not likely to be useful are dropped, and dummy variables are created with categorical data to enable some classifiers to perform better. The interval "days" may be useful for examining habits of adopted users over time, but it is not useful here for predicting future users as it is an expression of longevity with the service that may be intrinsically coupled with adoption.

With the remaining features examined via three classifiers, it seems difficult to predict adoption by users. The gradient boosting classifier performed the best with an AUC-ROC of 0.60, test accuracy of 0.827, and F1 score of 0.75, but accurate predictions for the target classes are highly imbalanced due to class imbalance in the data, so next we will re-examine with undersampling of the majority target class. While AUC-ROC stayed about the same after undersampling of the majority target class, the test accuracy and F1 scores went down a lot. Precision and recall, however, show more balance for the two target classes. Next we will look at results after tuning of hyperparameters with this classifier.

Hyperparameter tuning on the full training dataset did not seem to yield a better model in terms of multiple metrics. Likewise, with the criteria chosen, tuning of the model with undersampled target data did not improve predictive metrics, and results were similar to without tuning on with the undersampled data. Next, feature importance plots are shown with tuned data from the full dataset and the undersampled data.

The patterns shown in the feature importance plots indicate that organizational use of the service is strongly associated with adoption by users, as organization ID by far sweeps the list of features for predictive power here. The creation source and email domains also show some differences but to a much lesser degree. Interestingly, though invited members appeared to skew more toward adoption in the exploratory data analysis above, invited status itself seems not a strongly predictive feature. Undersampling the majority target affects relative importance of these lesser features, though the AUC-ROC was similar with or without undersampling, but with error differing for assignment to each target class.

Many of the features included here are worth inclusion in another analysis, but given how important organizational ID is, the sorts of features to explore in the future would include factors related to this. It is possible some organizations on the whole have adopted this service and not other competitor services. Examination of organization characteristics would likely be useful, and things to consider in the future may be: size of each organization, organization type, organization revenue, cost or levels of service chosen by the organization (if there are tiers), levels of interaction with the service between organizational members with each other versus people outside the organization, device types (if this is phone or computer-related), and user demographics, if available.