Relax Take Home Challenge Write Up

The goal of this challenge is to identify factors that influence user adoption. This is a classification problem, so I opted for Logistic Regression, Random Forest, and Gradient Boosting models. Support Vector Machines are also useful for classification problems but they are not known for their interpretability and were thus not used. At every stage, I tested logistic regression, random forest, and gradient boosting models but I am only discussing the logistic regression models below because they performed the best out of the three.

After running the baseline models, it became apparent that imbalance in class size caused the model to only predict the majority class in order to boost the accuracy score. As a result, I changed the model metric of choice to an F1 score and observed the confusion matrix. The baseline model had an F1 of 0.0, which indicates that the model was not able to adequately detect any classes.

My next attempt was to use various resampling methods to even out the class sizes. First, I upsampled the minority class to match the size of the majority class and fit the upsampled data to a logistic regression model. There was a definite improvement in the model's predictions, with a new F1 score of 0.25. Then, I tried using SMOTE (Synthetic Minority Over-sampling Technique) along with randomly undersampling from the majority class. Unfortunately, predictions using this data were just as bad as they were prior to applying resampling methods, yielding an F1 score of 0.0.

Finally, I used five-fold cross validation to tune the regularization parameter for the logistic regression model and fit the model with the first batch of resampled data (where only the minority class was resampled). This model was the best so far, with an F1 score of 0.25, so I extracted the coefficients of each feature in order to determine which ones were most important to the outcome.

The feature importance rankings from this model suggest that signing up through a Google account is the largest positive impact, and signing up through a personal project is the largest negative impact. The model will need further development in order to draw actionable results from it. My next steps, if I were tasked with this project, would be to tune the hyperparameters for the logistic regression model and potentially seek out other ways to resample the dataset.

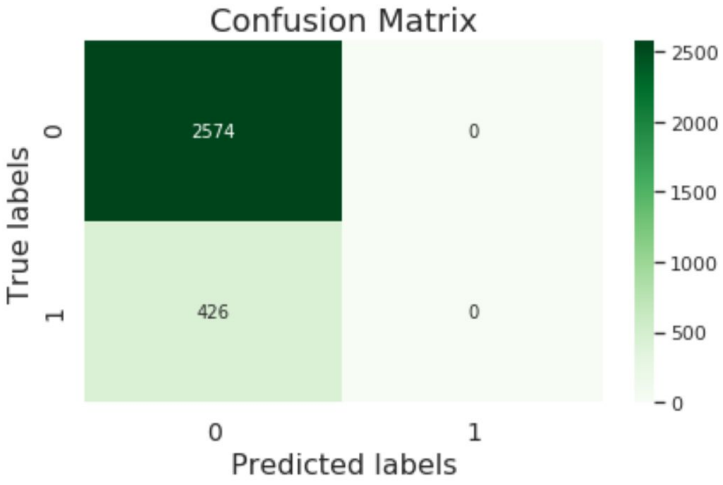

Figure 1 (top): Confusion matrix for baseline model. We can see that the value for the predicted label of 1 (i.e. how many times the model predicted that there would be user adoption) is zero.

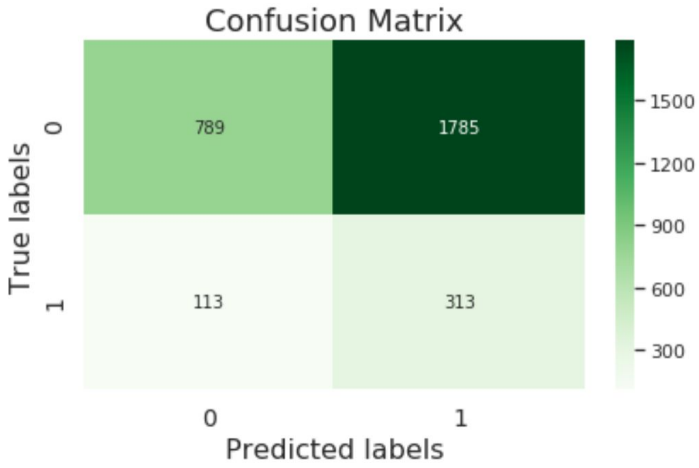

Figure 2 (middle): Confusion matrix for logistic regression model after upsampling minority class and performing cross validation. While these quantities are not ideal, they represent an improvement over the baseline.

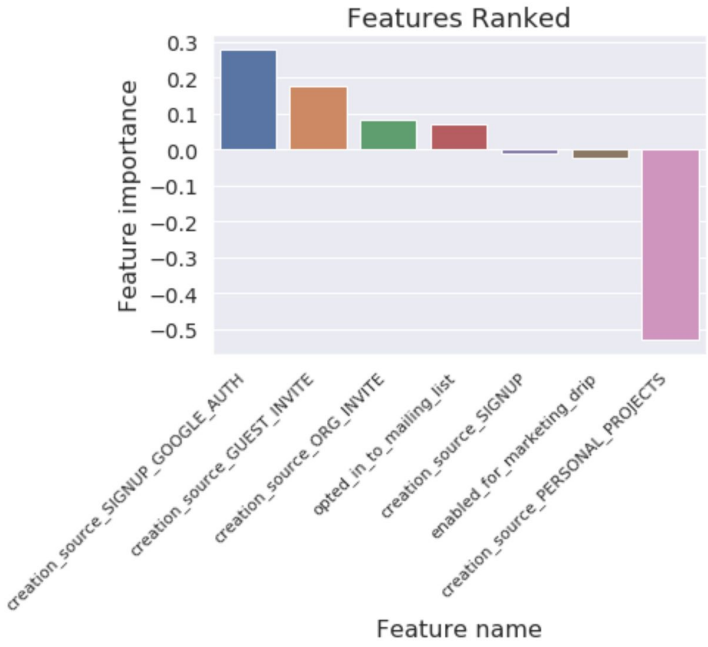

Figure 3 (bottom): Feature importance ranking for logistic regression model. The source of the user account's creation seems to have a large impact over user adoption. The model would need to be developed further to make reliable conclusions.