Relax Challenge Report

After first cleaning the data and making it manageable I tried using a Random forest classification technique on the data in order to find out that it is the organization that is the most predictive factor in being able to successfully predict whether a user would be an adopted user. However, I started to investigate this further by seeing if a separate model might perform better than Random Forest, and I found that Logistic regression did a better job in terms of accuracy. However, this is misleading because the recall was 50% which is terrible. I then printed a confusion matrix and I found that the models were always predicting zero, and since only 11% of users ended up being an adopted user, always predicting zero can give you an accuracy of 88% or so. Which is what I was finding, because it is right 88% of the time when always predicting zero.

Therefore, I found two resampling algorithms to combat the fact that there is unbalanced data. I utilized the class_weight = 'balanced_subsample' parameter of the Random Forest Classifier, and I found that the organization that the user was under was still a high contributor to the model. This time the model did predict some individuals being an adopted user. Table 1 shows the top 5 features in the dataset that contributed to the predictions.

The fact that it is the organization that leads to the ability of the model to correctly classify the data makes sense because if the organization is supporting this product then it would make it more likely that a user would log in more often. However, the classification of adopted user, is quite high in some sense. It appears since only 11% of 12,000 users were adopted users, that the product doesn't need to be logged into that often. This might be something that the company could look into to promote their product. If they would like to increase the amount of adopted users then either decrease the criteria, or make more of an incentive for users to log in on consecutive days.

Table 1: Feature selections from Random Forest Classifier

Feature	**Importance**
Organization	0.9138
Have a friend sign up	0.0206
Part of mailing list	0.016
Marketing Drip	0.0157
Created from personal projects	0.0146