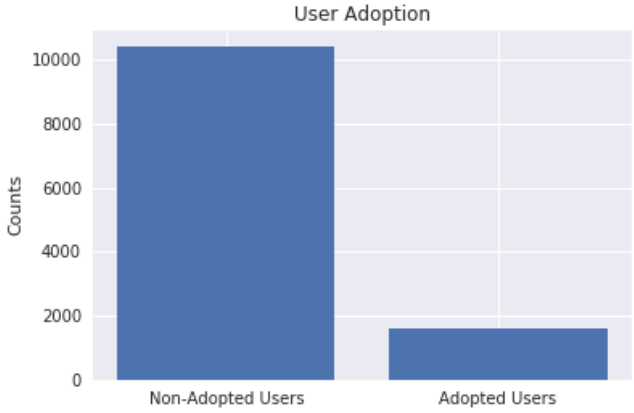
Relax Data Challenge Problem:

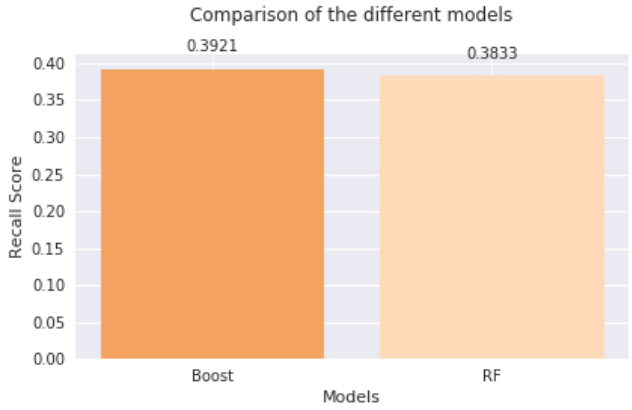
The Problem: Defining an "adopted user" as a user who has logged into the product on three separate days in at least one seven day period, identify which factors predict future user adoption.

EDA comments: 13% of users have adopted. Adopted Users from “Creation Source” are largest in “Org_Invite” by volume and “Personal_Projects” by percent. The overall trend is growth in signups and last session activity.



Model Selection and Model Evaluation: We picked a Random Forest and a Gradient Boosting model, both do well with imbalanced data and provide easy interpretability. Interpretability is key since we want to be able to understand the factors that predict future user adoption. The confusion matrix will give us a picture of where the model is making tradeoff between the class labels. The recall score will measure of how many relevant items are selected, that is how many positives are caught.

Modeling: The model was severely impacted by the imbalanced data, and initially did not predict any adopted users. The accuracy score at 87% was miss leading. Addressing the imbalance issue, we used a resampling method called SMOTE (Synthetic Minority Oversampling Technique) which uses a k-nearest neighbor model creating new instances of the minority class. Using the SMOTE, we achieved marked improvement of 39% in the recall score (178 predicted true positive labels).We compared the boosted model vs. a random forest model which only achieved 38%. We attempted to improve the Gradient Boost Model by tuning for recall score. In doing so, we saw the predicted true labels increased by 67 but we lost predicted true negatives 478. The recall score captured this gain but the accuracy score decreased. This is where the business needs should give guidance. If we solely care about predicting users who will adapt (predicted true only) then the tuned model is best. If we can about predicting for users who fail to adopt then the previous model (not tuning for recall) would be better. This highlights how we can adjust this model to target which prediction (or group) we care more about.



Last Remarks: We look at the features this model used in making its decision. We see ‘Personal Projects,’ ‘enabled marketing drip’ and ‘original invite' are leading features that our model used. This information is valuable and gives insight into what could be drivers for drawing more adopted users. We see that Org_Invite and Personal_Projects are two good candidates. A marketing push highlighting the platform’s use for personal projects could draw new users who are more likely to be adopted users. We can also look for ways to transition ‘Guest_Invites’ accounts to ones with full access.