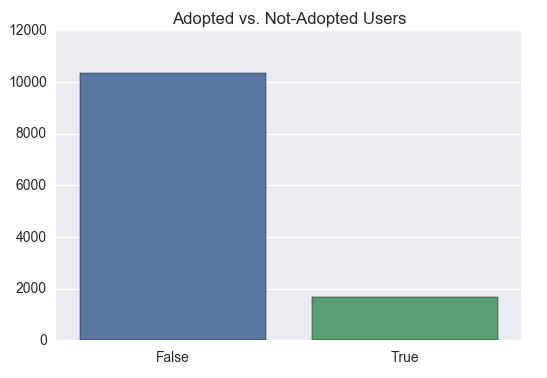
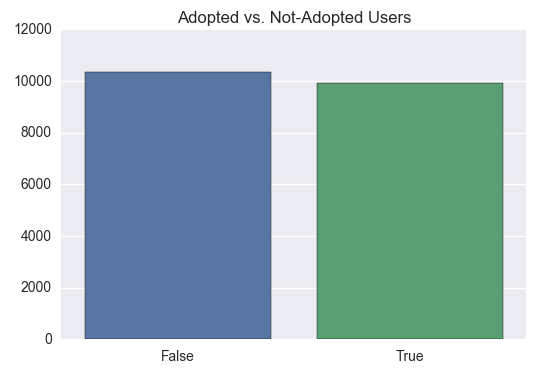
In order to evaluate which factors predict future user adoption, I began by reading in both datasets and assessing data quality. Starting with 'takehome\_user\_engagement.csv’, the log dataset, I had to re-encode the data in utf-8 so as to open it properly. Then I counted the number of unique user ID numbers at 8,823 users, which represents the number of users who have logged onto the application. I then read in ‘takehome\_users.csv’, the users dataset, and inspected it to find that there are 12,000 unique user ID numbers where 3,177 users do not show ever having logged onto the application, which explains what seemed to be a discrepancy between the datasets.

I then needed to use the log dataset in order to determine which users were considered ‘adopted users’ for use as the target class in the users dataset. After creating the labels for each user, I found that approximately 18.77% of users who have used the application are considered adopted users. Further, approximately 13.80% of all users who have ever signed up for the application are considered adopted users. We can see the latter distribution of the target class through the bar graph on the right.

In order to adjust for this and ensure an appropriate model, I employed the SMOTE algorithm to synthetically enhance the True, minority class. After adjusting for this, the distribution of the target class is shown by the new bar graph on the right. I then dropped any features that were not predictive in nature and was left with: creation\_source, opted\_in\_to\_mailing\_list, enabled\_for\_marketing\_drip, org\_id, and the target class.

Using the four remaining features, I decided upon the random forest classifier (with 10 decision trees) for modeling. My decision was such due to increases in model appropriateness over decision tree, k-nearest neighbors, logistic regression, and bagged decision trees. I would have favored a decision tree for its increased interpretability, but its accuracy was ~10% lower than that of the random forest classifier. Testing the random forest classifier, we obtain ~81.66% accuracy on the test set with a True TPR of 88.27%, False TPR of 74.77%, True FPR of 25.23%, and False FPR of 11.73%. We see that balancing the target class has helped in creating a valid model.

The resulting model determined that the organization ID was the most predictive feature for determining future user adoption, and this was followed by whether a user opted in to the marketing email list, whether a user was enabled for marketing email drip, whether the user was invited to join another user’s personal workspace as a way of account creation, and then the remaining four avenues of account creation.