Relax Take Home Challenge Report

By: Seung Chi

Data Cleaning and Munging

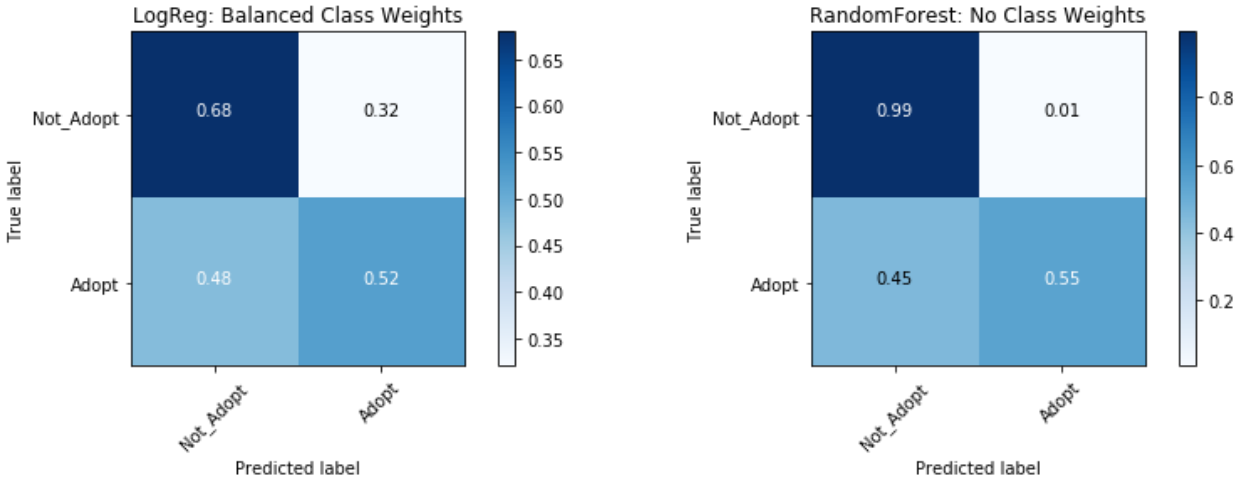
I started with the takehome_user_engagement.csv file to identify adopted users based on the criteria of at least 3 logins within a 7 day time window. This information was added as a 0/1 column to the takehome_users.csv file, as the target variable for predictive modeling. The data contained null values in the last_session_creation_time and invited_by_user_id columns. I identified the distribution to be heavily skewed to the right for the last_session_creation_time column, so imputing the mode value made most sense. For the invited_by_user_id, the nan values mostly likely means the user was not invited, so the nan values were converted to 0, while others were converted to 1 to indicate invitation.

I scaled down the last_session_creation_time value by 10^9, converted the creation_time to days since the max value, split the email to just the domain names, dropped the name column, and converted the org_id to categorical. Then, I used get_dummies to convert the categorical columns (email, creation_source, org_id) for the machine learning step.

Machine Learning

Since this is a supervised classification problem, I used logistic regression and random forest classifier. The logistic regression performed poorly just predicting all not_adopt for the users since the original data is imbalanced with 13% adopted users. With class weights parameter set to balanced, the prediction became more balanced for correct user adopted labeling. However, overall, the performance is not good for the logistic regression.

The random forest classifier performs much better with 55% true predictions for adoption rate while maintaining total accuracy. Interestingly, the class weight parameter set to balanced does little to the model's performance, in contrast to the logistic regression models.



Factors that Predict Future User Adoption

From the feature importances for the random forest classifier, the top 5 factors are:

- Last_session_creation_time: 0.3627
- Creation_time: 0.1158
- Creation_source_google: 0.0178
- Opted_in_to_mailing_list: 0.0117
- creation_source_Signup: 0.0115

Considerations and Further Research

More can be done on the RandomForestClassifier with hyperparameter tuning. Also, it would be interesting to note if the attributes of the inviting user id plays a role in determining adoption rate.