Relax Challenge

See jupyter notebooks in github folder for more details.

Feature Selection

The features selected for model construction were as follows:

- ‘opted_in_to_mailing_list' : boolean indicating whether they opted into receiving marketing emails
- 'enabled_for_marketing_drip' : boolean indicating whether they are on regular marketing email drip
- 'creation_source_GUEST_INVITE' : boolean indicating if account was created by being invited to an organization as a guest (limited permissions)
- 'creation_source_ORG_INVITE' :boolean indicating if account was created by being invited to an organization as a full member
- 'creation_source_PERSONAL_PROJECTS' : boolean indicating if account was created by being invited to another user’s personal workspace
- 'creation_source_SIGNUP' : boolean indicating if account was created by being signing up on website
- 'creation_source_SIGNUP_GOOGLE_AUTH’ : boolean indicating if account was created using Google Authentication
- 'invited' : boolean indicating if a user invited them
- 'login_after_signup' : boolean indicating if user signed into account after signup

Logistic Regression

The first model I constructed was a logistic regression model. Unfortunately the model returns all 0 values, meaning it predicts that no users will be retained.

The confusion matrix for the Logistic Regression model:

2635	0
365	0

The coefficients for the model were:

feature	coefficient
opted_in_to_mailing_list	0.083662
enabled_for_marketing_drip	0.027309
creation_source_GUEST_INVITE	0.217995
creation_source_ORG_INVITE	-0.157148
creation_source_PERSONAL_PROJECTS	0.179136
creation_source_SIGNUP	-0.103685
creation_source_SIGNUP_GOOGLE_AUTH	-0.136518
invited	0.060847
login_after_signup	4.701521

Although the model was not helpful to determine which users are retained, it can be useful to see that login_after_signup is the most important feature.

K Neighbors Classifier

The next model I built was a K Neighbors Classifier. The model proved to be better at predicting retention, although the overall accuracy was lower due to the skewed nature of the data.
The confusion matrix for the K Neighbors model:

2142	493
271	94

The K Neighbors model performed better but does not provide information on which features are the most important for user retention.