RELAX INC - CASE STUDY
A supporting Jupyter notebook is in this same folder (Relax_Inc.ipynb).

1 DATA & FEATURE ENGINEERING
I loaded the data as provided in the CSV files. For the user engagement data, I flagged each user id as adopted or not if the user logged in for 3 days or more in any 7 day window. For the user attribute (feature) data, I did the following:

- reviewed null values (data are complete, any nulls are equivalent to zero values)
- removed duplicates using email as a key (retaining record with first creation_time)
- one-hot encoded column "creation_source"
- binned org_id by count of each distinct org_id
- converted invited_by_user_id to an indicator that flags invited by another user or not

The resulting data set has 9 features (opted_in_to_mailing_list, enabled_for_marketing_drip, guest_inv_flg, org_inv_flg, per_proj_flg, signup_flg, signup_googauth_flg, org_id_idx, invited_flg) and a binary target variable (adopt_flg) with value 1 indicating product adoption and 0 indicating otherwise. All of the features are binary except for org_id_idx which has 4 numerical categorical values.

2 RESULTS
As an initial exploratory analysis, for each feature I calculated the mean of adopt_flg over the feature category values. The mean of adopt_flg is equivalent to the probability (percent) of product adoption within each category value. This analysis suggests that some variables have no average effect on adoption. For example, percent adoption is virtually the same regardless of mailing list opt in or opt out. Based on this initial analysis, I reduced the feature set to 6 features: guest_inv_flg, per_proj_flg, signup_flg, signup_googauth_flg, org_id_idx, invited_flg.

Using these 6 features, I fit a logistic regression model and evaluated it by cross validation on a training data set and assessed the model on a separate test data set. The accuracy of the logistic regression model on the training data set is 87% (0.5% st dev) and 88% on the test data. Using feature importance, the below chart shows the relative importance of the features as factors that predict future user adoption.