Relax Take-home Challenge

By: Vinh Mai

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.

# Data Wrangling & Preprocessing

## ‘takehome\_users.csv’ contains:

* Four object columns
  + ‘creation\_time’, ‘name’, ‘email’, ‘creation\_source’.
  + Dropped ‘name’ and ‘email’ as nothing can be predicted from these columns.
  + ‘creation\_time’ was converted to datetime format. Only the year was used in the modeling process because month, day and time was insignificant in the predictive modeling.
  + ‘creation\_source’ was one-hot encoded.
* Six numeric columns
  + ‘object\_id’, ‘last\_session\_creation\_time’, ‘opted\_in\_to\_mailing\_list’, ‘enabled\_for\_marketing\_drip’, ‘org\_id’, ‘invited\_by\_use\_id’.
  + ‘object\_id’ was made the index.
  + Using ‘org\_id’ for reference, a binary column was created called ‘large\_org’ (where ‘1’ is an orge consisting of 29 or more users, and ‘0’ for everything less.)
  + ‘invited\_by\_user\_id’ had null values for anyone not invited by a user. To simplify, a column was added called ‘user\_invited’, where any null values were changed to ‘0’ and anyone with an invite was changed to '1’.
  + ‘last\_session\_creation\_time’ was not relevant and therefore dropped.

1597 users logged on at least three times in a seven-day period.

# EDA

Using a correlation matrix to determine feature relationships, it was found that the feature with the best connection to user adoption was the ‘creation\_source’, where the invitation was “GUEST\_INVITE”.

# Modeling

## The following models was tested: Decision Tree

* RandomForestClassifier
* ExtraTreesClassifier
* GradientBoostingClassifier

Since most of the features were made binary/Boolean, the choice to use a decision tree was made.

## Cross-Validation

* RandomForestClassifier ROC\_AUC score = 0.6398
* ExtraTreesClassifier ROC\_AUC score = 0.6399
* GradientBoostingClassifier ROC\_AUC score = 0.6376

ExtraTreesClassifier had the best ROC\_AUC score of the tested models, with an ROC\_AUC score of 0. 6399and an accuracy score of 0.5965. The best parameters for the model are 'bootstrap': True, 'max\_depth': 8, 'max\_samples': 0.975, 'n\_estimators': 50. The feature ‘creation\_year’ has the highest correlation to user adoption.