Relax Take Home Datasets

The goal of this project is to determine which factors predict whether or not a user will become an adopter. Two datasets were provided. One contained timestamps of user logins, and the other contained information about specific users. There are a total of 12000 users; of those, 8823 users logged into the service at some point.

We set a definition of an adopted user, that is, a user who has logged in at least three times in a week is considered to be an “adopter”. The determination of who is an adopter is done with a function which parses lists of days (integers) and determines whether 3 appear within a 7 day period. Once our data has a binary label, we plan to use supervised learning with a binary classification model.

The data is unbalance, with the majority class outnumbering the minority 7:1:
Counter({0.0: 10294, 1.0: 1556}). Unfortunately, this does result in a clever support vector machine determining that *all* users are not adopters.

We can weight the classes using the sklearn tool computer_class_weight and assign those values to the instantiation of the learning models. In this case we tried Random Forest Classifier and Linear Support Vector Classifier and both gave unsatisfactory results, which did not improve upon the initial hypothesis of, “All users are not adopters”.

```
              precision    recall  f1-score   support

         0.0       0.88      0.70      0.78      1560
         1.0       0.13      0.31      0.18       218

accuracy                       0.65      1778
 macro avg         0.50      0.51      0.48      1778
weighted avg       0.79      0.65      0.71      1778
```

The current accuracy is: 0.618, which is much worse than merely assigning everything to the majority class.

There are improvements I would suggest to the feature engineering for this model. We used categoricals for most features and pandas’ built-in tool to encode dummies. We dropped email, which could be used by extracting the domain and changing it to a categorical. Additionally, it could be very useful to predict adoption using a feature consisting of the length of time between the creation of the account and the first login. I would hypothesize that the longer someone takes to actually log into the system, the less likely they are to be an adopter. This feature could be created from the existing features.

.