Relax Data Science Challenge Report

Corey Wade

September 1, 2018

Preliminary data analysis does not reveal any smoking guns for predicting future adopted_users. Each column of data had fairly low correlations with respect to adopted_users.

An interesting column that needed to be factored is the “creation_source.” Summing the values for all users and adopted users, it appears that guest invites are more likely to become adopted users, personal projects are less likely to become adopted users, and signup with Google Authorization is a little more likely to become an adopted user.

```
# Show distribution of creation_source for all users
df_users[‘creation_source’].value_counts(normalize=True)
```

ORG_INVITE	0.354500
GUEST_INVITE	0.180250
PERSONAL_PROJECTS	0.175917
SIGNUP	0.173917
SIGNUP_GOOGLE_AUTH	0.115417

```
# Show distribution of creation source for adopted users only
df_adopted_users[‘creation_source'].value_counts(normalize=True)
```

ORG_INVITE	0.346618
GUEST_INVITE	0.222826
SIGNUP	0.182367
SIGNUP_GOOGLE_AUTH	0.144324
PERSONAL_PROJECTS	0.103865

I considered using time columns as data, but discarded the idea because the requirements for an adopted user are time based.

I created a new column, invited_by_user_id, that returned a 1 or 0 depending on whether a user had been invited or not.

My first attempt at running machine learning tests had high 86% accuracy, but the classification report revealed that they were predicting 100% non-adopted users. The data revealed that only 14% of all users are adopted users.

The data was oversampled to get more accurate learning results. A decision tree ({'criterion': 'gini', 'max_depth': None, 'max_features': 3, 'min_samples_leaf': 1} produced the best results with 83% accuracy, considerably outperforming Naive Bayes, Logistic Regression and Random Forests which obtained approximately 55% accuracy.

In conclusion, I would recommend collecting more data from each user. This would create more columns, and likely improve results.

All code, comments, reports and files are included in the attached Jupyter Notebook.