Predicting Pull Request Acceptance on GitHub from Social Factors

Matthew Heston

Hypothesis

- Social factors can be used to accurately predict whether or not a pull request from a first time contributor will be accepted by project maintainers on GitHub.
  - Community members who have been active in previous discussions are more likely to have their code changes accepted.
  - Community members who have other popular projects are more likely to have their code changes accepted.
  - Pull requests that generate a lot of comments are more likely to have their code changes accepted.

Data

- 45 Repositories
- Closed requests only
- First pull requests
    - 13,383 total
    - 4,352 merged (32.5%)
    - 9,031 not merged (67.5%)

Features

- Number of pull requests commented on
  - Merged
    - Min: 0
    - Max: 900
    - Mean: 0.892
    - Median: 0
    - SD: 16.91

- Number of pull requests commented on
  - Not merged
    - Min: 0
    - Max: 173
    - Mean: 0.245
    - Median: 0
    - SD: 3.18

Features

- Number of stars for selected repos
    - Merged
        - Min: 0
        - Max: 27024
        - Mean: 133.51
        - Median: 5
        - SD: 995.42

- Number of stars for selected repos
    - Not merged
        - Min: 0
        - Max: 20709
        - Mean: 85.95
        - Median: 2
        - SD: 532.11

Features

- Number of comments on pull request
  - Merged
    - Min: 0
    - Max: 200
    - Mean: 2.239
    - Median: 1
    - SD: 5.18

- Number of comments on pull request
  - Not merged
    - Min: 0
    - Max: 85
    - Mean: 3.437
    - Median: 2
    - SD: 4.77

Experiments: Logistic Regression

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.955e-01  2.393e-02 -20.707  < 2e-16 ***
past_comments   2.230e-02  2.753e-03   8.100  5.5e-16 ***
popular_repos   6.590e-05  2.912e-05   2.263   0.0236 *
comments_on_pr -9.286e-02  6.282e-03 -14.782  < 2e-16 ***

Null deviance: 16882  on 13382  degrees of freedom
Residual deviance: 16569  on 13379  degrees of freedom
```

Experiments: Logistic Regression

```
                      OR     2.5 %    97.5 %
(Intercept)    0.6092618 0.5813450 0.6385208
past_comments  1.0225516 1.0174156 1.0306517
popular_repos  1.0000659 1.0000109 1.0001261
comments_on_pr 0.9113242 0.9000067 0.9224457
```

Experiments: Classifiers

	Accuracy	F1
Logistic Regression	60.2	51.5
Naive Bayes	66.9	12.9
Decision Tree	70.9	44.9

trained on number of previous pull requests commented on and number of comments on current pull request
using 10 fold cross validation

Experiments: Classifiers

	Accuracy	F1
Logistic Regression	60.4	52.9
Naive Bayes	35.2	38.1
Decision Tree	68.6	44.1

trained on number of previous pull requests commented on, number of comments on current pull request, number of commits, number of changes
using 10 fold cross validation

Experiments: Classifiers

	Accuracy	F1
Logistic Regression	50.0	48.1
Naive Bayes	66.1	2.6
Decision Tree	65.4	10.2

trained on number of number of commits and number of changes using 10 fold cross validation

Discussion

- Von Krogh, Georg, Sebastian Spaeth, and Karim R. Lakhani. "Community, joining, and specialization in open source software innovation: a case study."
  - “Participants behaving according to a joining script (level and type of activity) are more likely to be granted access to the developer community than those participants that do not follow the project’s joining script.”
  - specialization and contribution barriers

Discussion

- Bryant, Susan L., Andrea Forte, and Amy Bruckman. "Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia."
  - “According to LPP, newcomers become members of a community initially by participating in peripheral yet productive tasks that contribute to the overall goal of the community”

Discussion

- Negative relationship between number of comments on pull request and acceptance
  - Complexity of task
  - Amount of time open

Future Work

- Prediction results not that good
- What other factors are important?
- Does the GitHub interface remove barriers to entry?
- What happens after the first pull request?
- Does each repository have their own joining script?
    - Does it change over time?