

# Appendix

## ECS 260 Project <sup>1</sup>

ARJUN BHARADWAJ AND CHRISTOPHER LOCK

Friday, December 2, 2016

<sup>1</sup><https://github.com/prudentprogrammer/Pull-Requests-Determinants>



# List of Figures

1	Result of Trees Selection - Feature Importance . . . . .	iv
2	Heatmap examining the correlations of features . . . . .	v
3	Histogram of Features . . . . .	vi
4	Histogram (Logged) of Features . . . . .	vii
5	AUC Curves for Models . . . . .	viii
6	Box Plot of Description Lengths: Merged vs Non-merged . . . . .	ix
7	Box Plot of Churn: Merged vs Non-merged . . . . .	ix
8	Box Plot of Additions: Merged vs Non-merged . . . . .	x
9	Box Plot of Deletions: Merged vs Non-merged . . . . .	x
10	Spree Plot for PCA . . . . .	xi
11	Biplot for PCA . . . . .	xi
12	Top 15 Feature Based on Univariate Selection . . . . .	xii
13	Top Features based on Ranking for Recursive Feature . . . . .	xii
14	Pull Request Acceptance for each month over period of 6 months . . . . .	xiii

Figure 1: Result of Trees Selection - Feature Importance

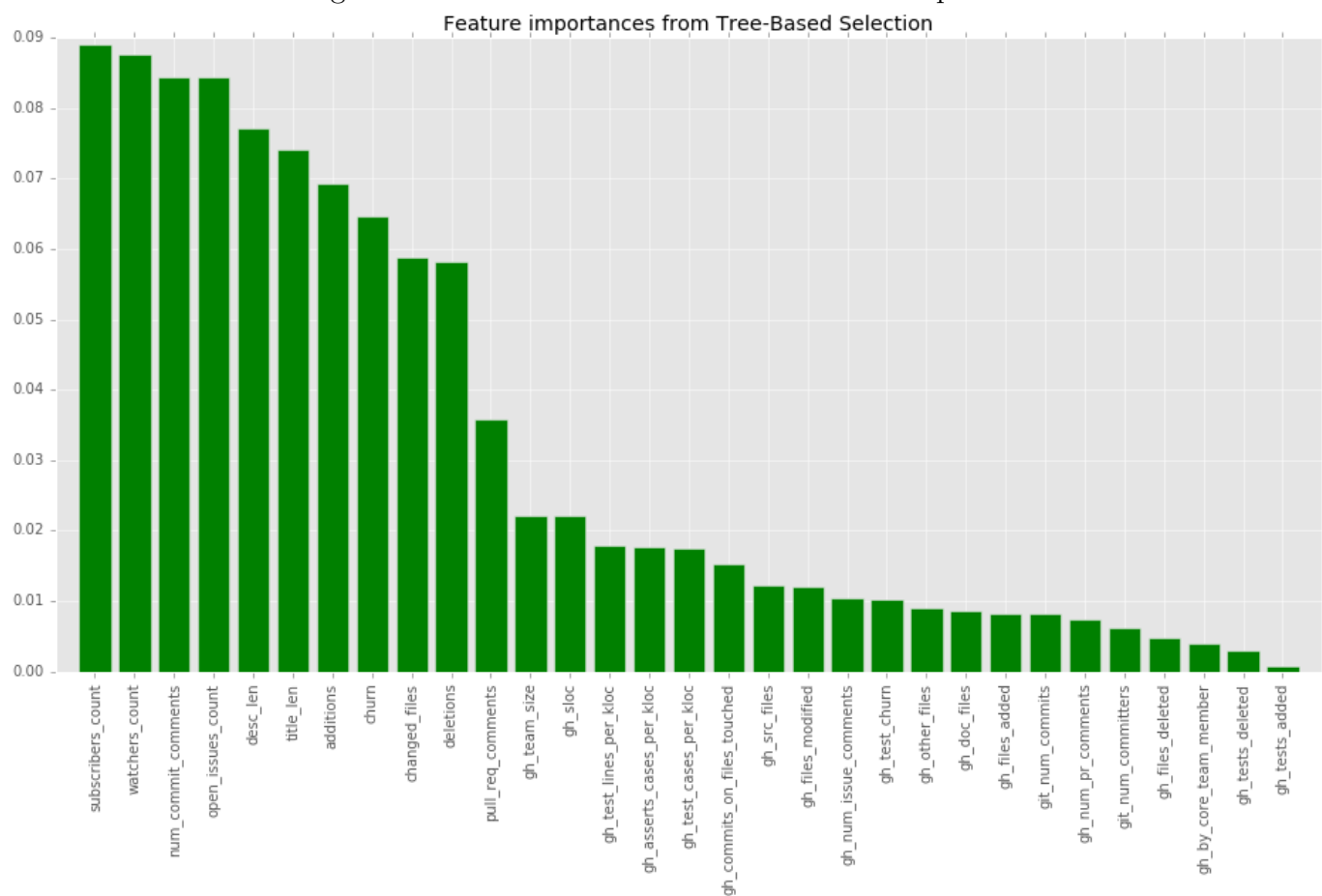


Figure 2: Heatmap examining the correlations of features

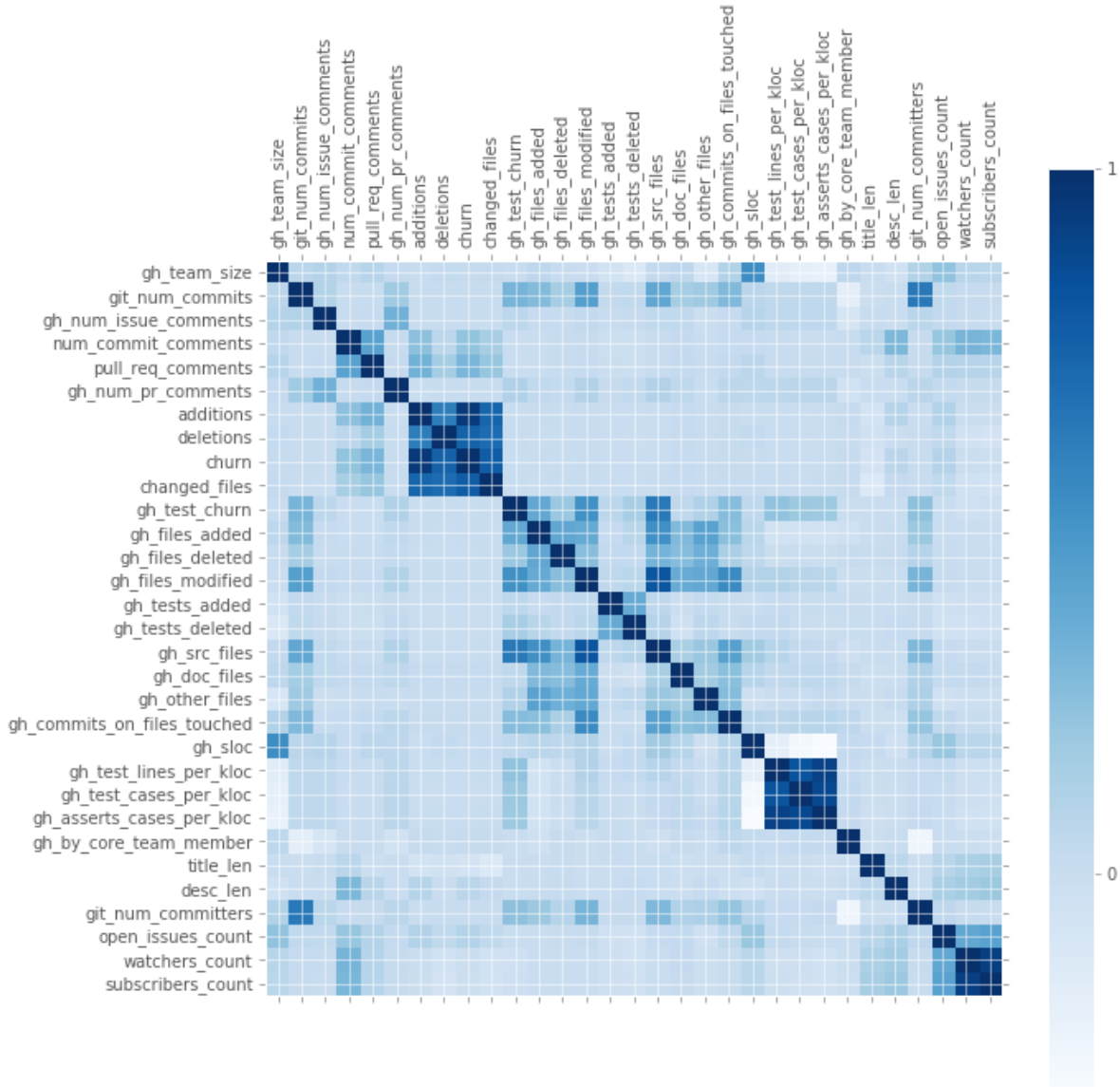


Figure 3: Histogram of Features



Figure 4: Histogram (Logged) of Features

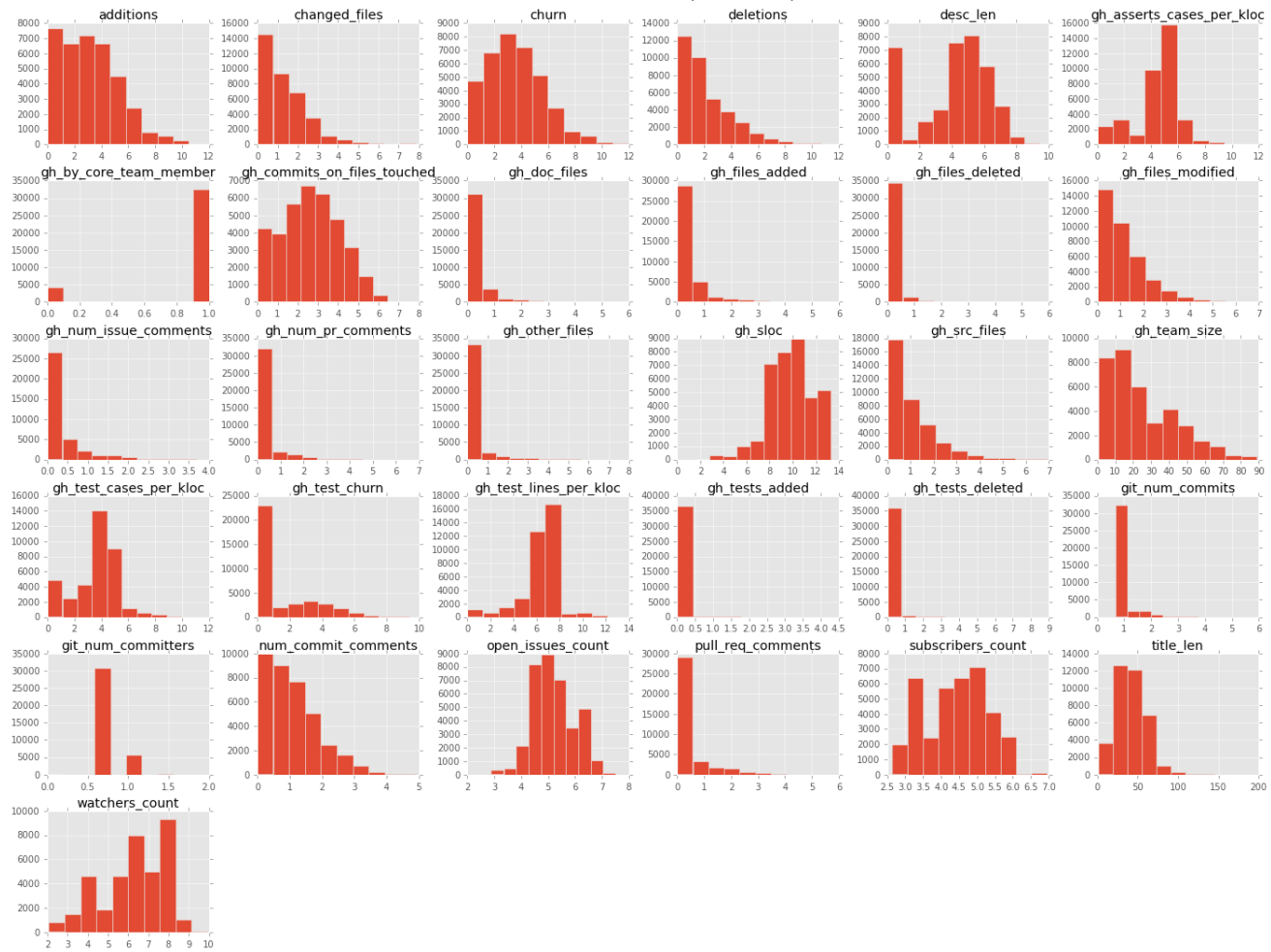


Figure 5: AUC Curves for Models

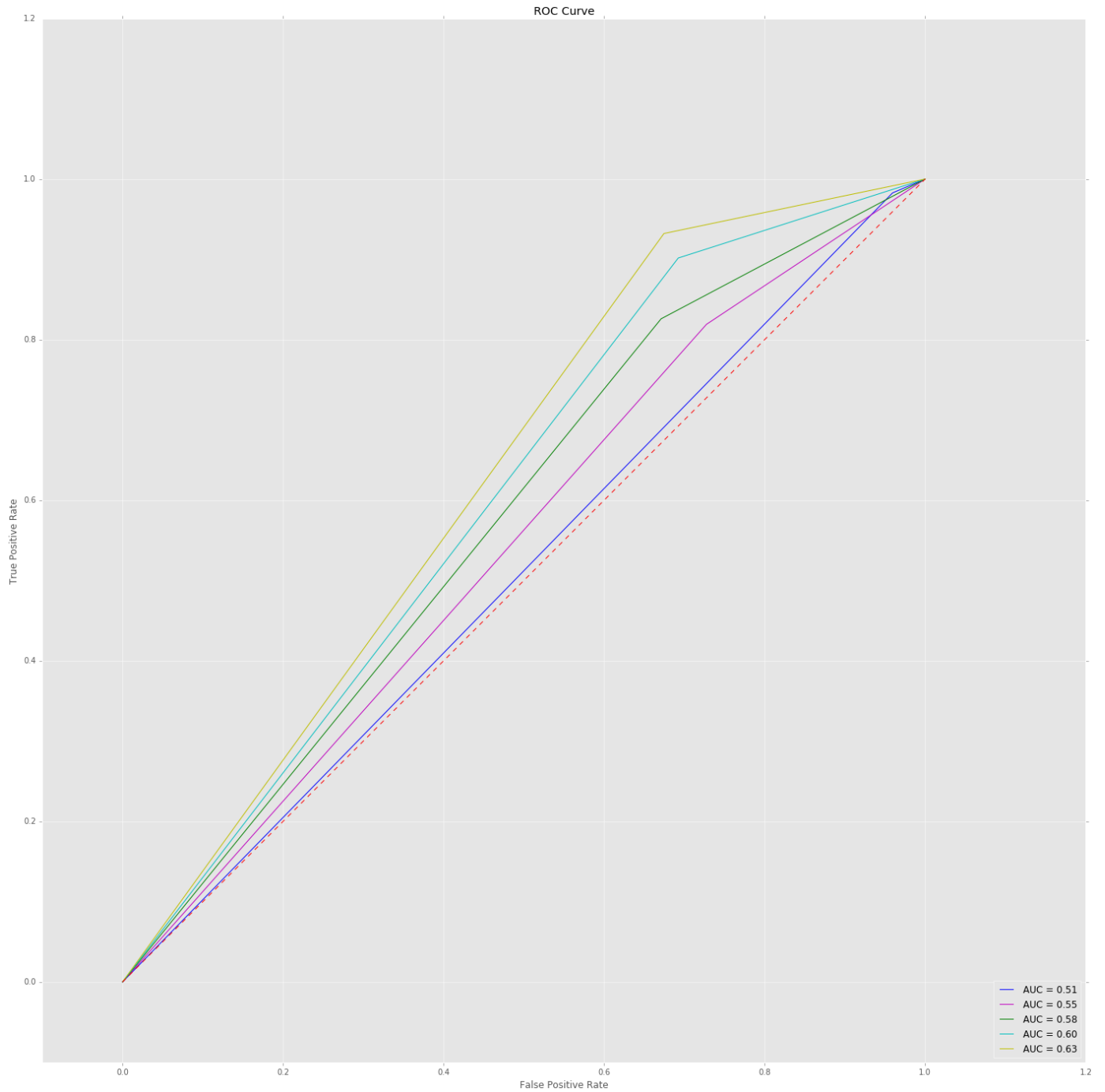




Figure 6: Box Plot of Description Lengths: Merged vs Non-merged

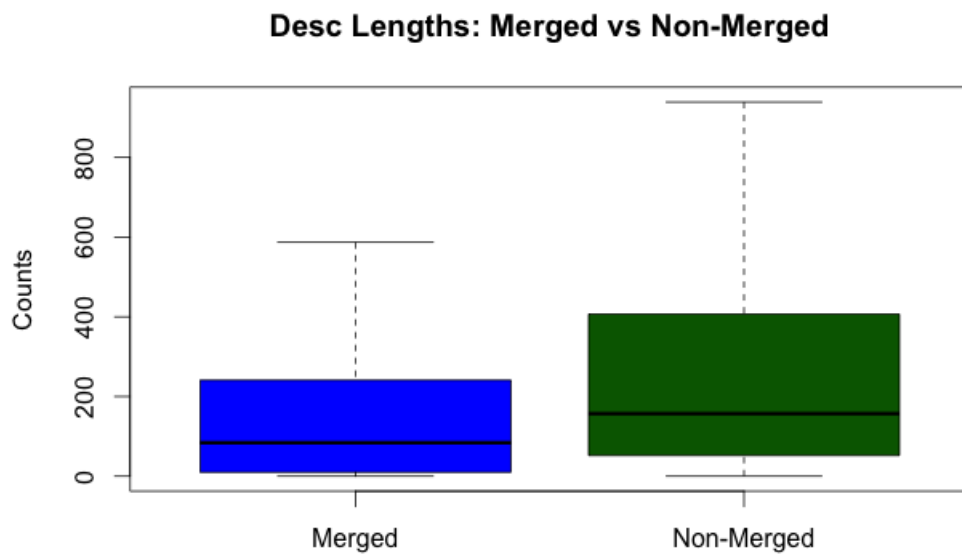


Figure 7: Box Plot of Churn: Merged vs Non-merged

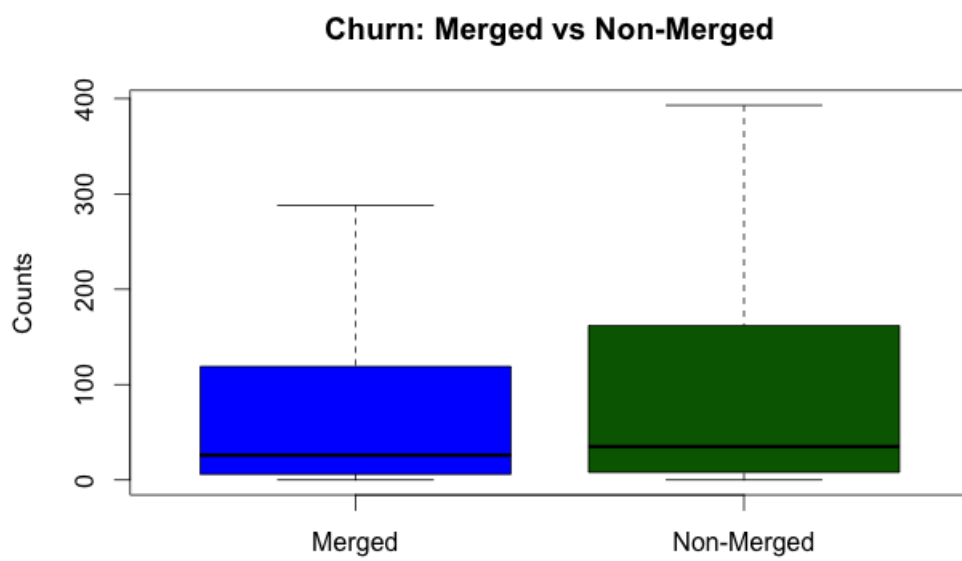


Figure 8: Box Plot of Additions: Merged vs Non-merged

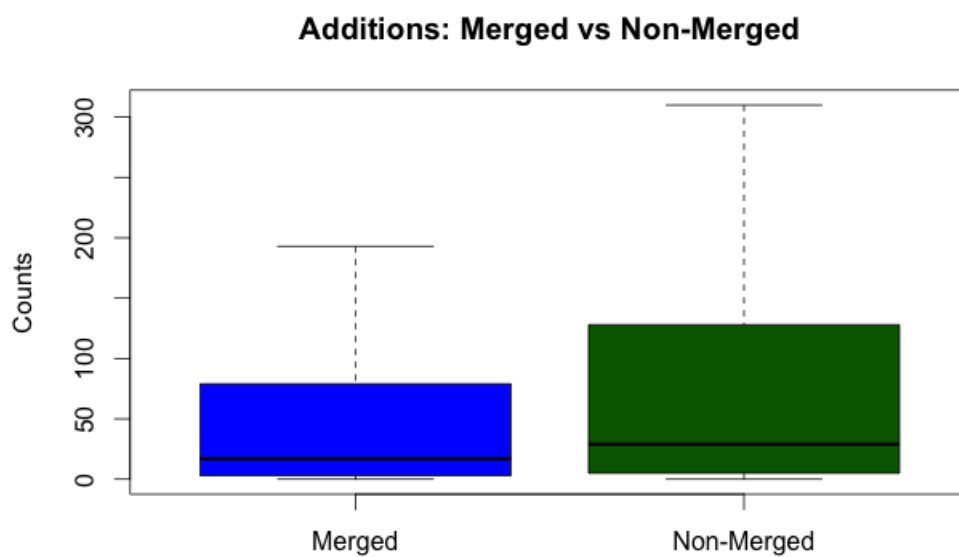


Figure 9: Box Plot of Deletions: Merged vs Non-merged

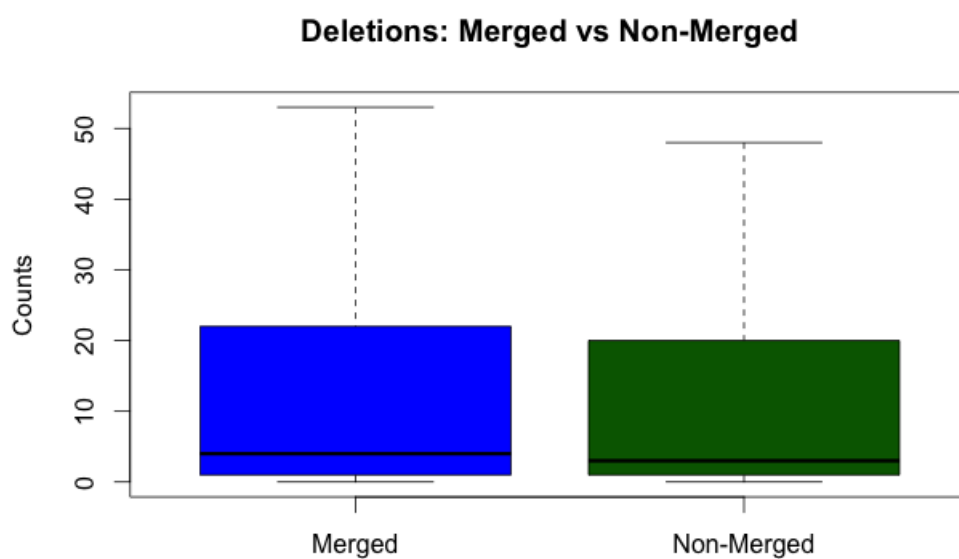


Figure 10: Spree Plot for PCA

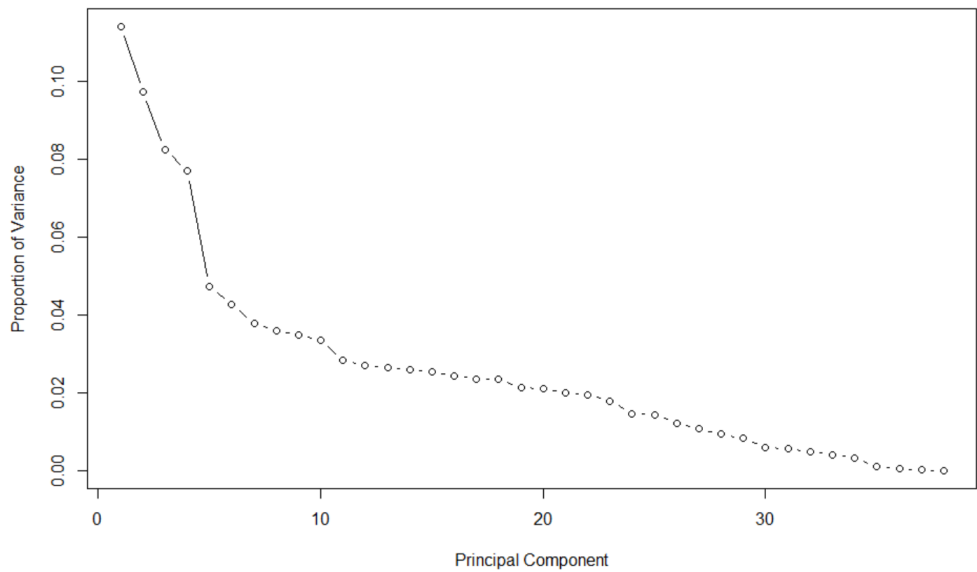


Figure 11: Biplot for PCA

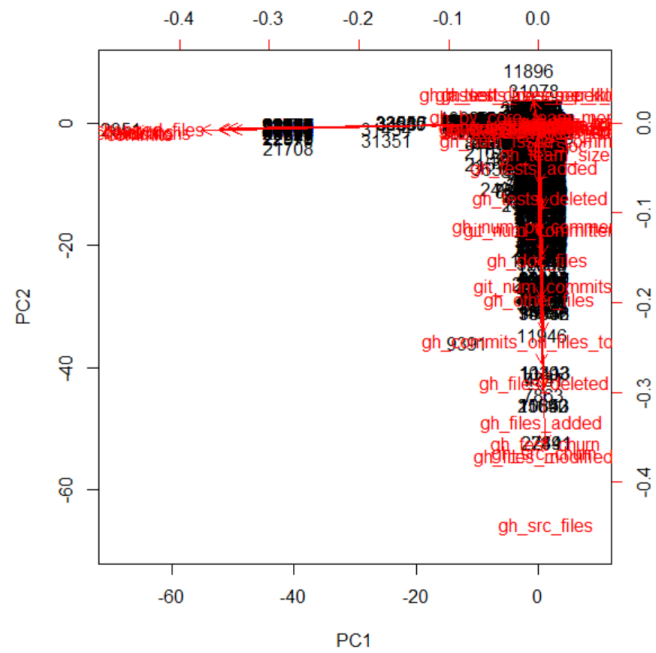


Figure 12: Top 15 Feature Based on Univariate Selection

## Univariate Selection - Chi<sup>2</sup> Selection

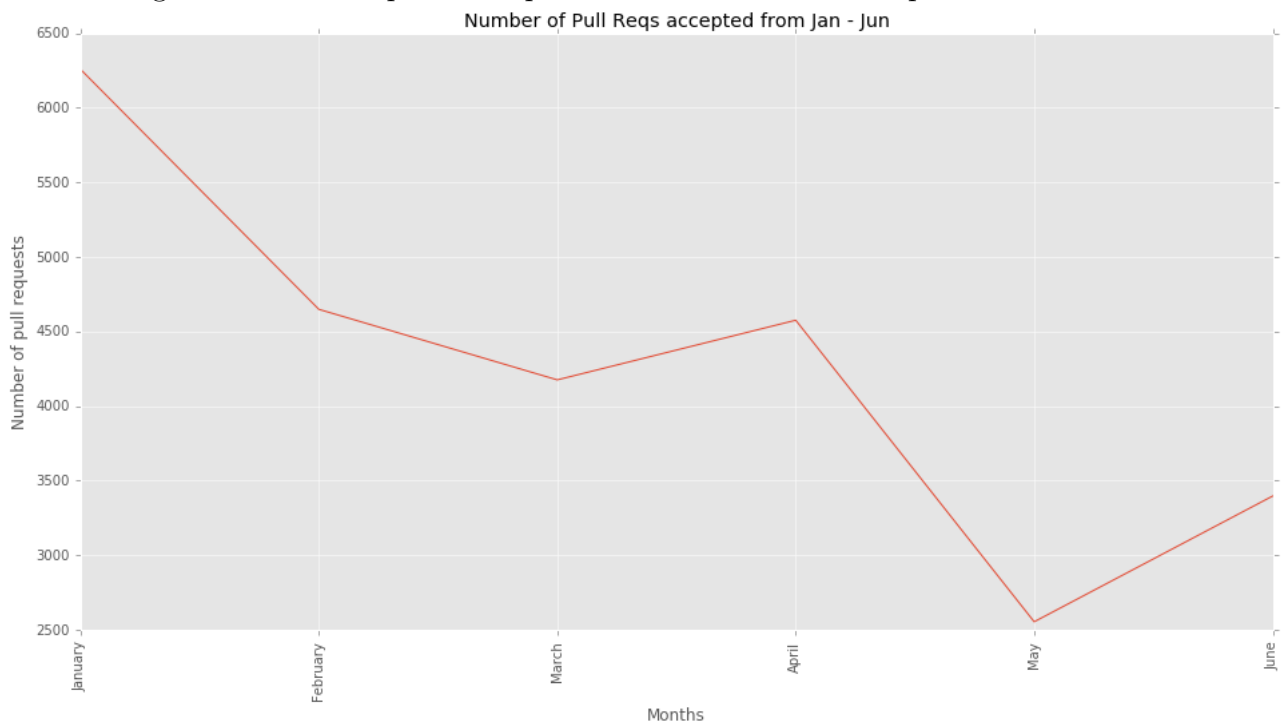
- |                      |                                 |
|----------------------|---------------------------------|
| 1) churn             | 9) subscribers_count            |
| 2) additions         | 10) gh_test_lines_per_kloc      |
| 3) gh_sloc           | 11) num_commit_comments         |
| 4) deletions         | 12) gh_test_churn               |
| 5) watchers_count    | 13) gh_commits_on_files_touched |
| 6) desc_len          | 14) title_len                   |
| 7) changed_files     | 15) gh_doc_files                |
| 8) open_issues_count |                                 |

Figure 13: Top Features based on Ranking for Recursive Feature

## Recursive Feature Elimination

- |                               |                                |
|-------------------------------|--------------------------------|
| ('gh_team_size', 1),          | ('gh_doc_files', 1),           |
| ('git_num_commits', 1),       | ('gh_by_core_team_member', 1), |
| ('gh_num_issue_comments', 1), | ('title_len', 1),              |
| ('num_commit_comments', 1),   | ('git_num_committers', 1),     |
| ('pull_req_comments', 1),     | ('subscribers_count', 2),      |
| ('gh_tests_added', 1),        | ('changed_files', 3),          |
|                               | ('gh_files_modified', 4),      |
|                               | ('gh_src_files', 5),           |

Figure 14: Pull Request Acceptance for each month over period of 6 months





# List of Tables

1	Accuracies of the Machine Learning Models . . . . .	xv
2	Cohen's Values (Effect size) for Merged vs Non-merged . . . . .	xv

Table 1: Accuracies of the Machine Learning Models

Logistic Regression	67.07%
$k$ NN	67.55%
Random Forests	71.81%
Adaboost	71.98%
SVM count	66.36%

Table 2: Cohen's Values (Effect size) for Merged vs Non-merged

Feature	Cohen Value	Effect size
Number of Commits	0.03	Very Small
Title Length	0.10	Small
Description Length	0.23	Small
Churn	0.22	Small
Additions	0.26	Small
Deletions	0.14	Small

## 0.1 Explanations (If necessary)

### 0.1.1 Explanation of Figure 1

In the figure, all the values add up to 1 and hence each feature's importance is normalized. The number of comments is the most discriminative feature in the dataset based on the result of 1,000 trees. However, the only downside related to random forests is that if two or more features are highly correlated, one feature may be ranked very highly while the information of the other feature(s) may not be fully captured. However since we are concerned with the predictive performance of a model collectively with all the features rather than examining individual ones, it should not be a problem under this context.

### 0.1.2 Explanation of Figure 2

Image of heatmap between pairwise features. Dark colors indicate strong correlation between variables.

### 0.1.3 Explanation of Figure 3

Distribution of features. As can be noticed they are skewed.

### 0.1.4 Explanation of Figure 4

Distribution of features, which are logged, to reduce the skewness.

### 0.1.5 Explanation of Figure 5

AUC Curves for various models.

### 0.1.6 Explanation of Figure 6,7,8,9

Box plots for various features for cases of merged vs nonmerged pull requests.

### 0.1.7 Explanation of Figure 10

The Spree plot displays principal components versus their corresponding eigenvalues, or in this context, their corresponding variance. A couple conclusions can be made from the spree plot: the first four principals have similar variances, ranging from 8% – 11%, that are each significantly higher than any other principal component's variance. These are the most meaningful principal components with PC1 accounting for 11% variance. The second observation is that 98% of the variance is accumulated within the first 30 principal components which implies we could comfortably model the data on 30 dimensions or features instead of the 37 tested.

### 0.1.8 Explanation of Figure 11

A biplot graphs all of the Principal components and their geometric relationships to each other. The biplot suggests that the `commits`, `additions`, `deletions`, and `churn` features contributed the most variance to PC1; the `gh_src_files` feature contributed the most variance to PC2(followed by `gh_src_churn`, `gh_test_churn`, `ghf_files_added`, and `gh_files_modified`); and the `subscribers_count`, `watchers_count`, `network_count`, `gh_test_lines_per_kloc`, `gh_test_cases_per_kloc`, and `gh_assets_cases_per_kloc` contributed the most variance to both PC3 and PC4. (Note: we found these conclusions by examining the actual used to generate the biplot, the biplot just provides an easy visual representation).

### 0.1.9 Explanation of Figure 12, 13

Results of other feature selection algorithms. Univariate selection by top features and recursive feature elimination by feature ranking (i.e 1 corresponds to top features etc.)

### 0.1.10 Explanation of Figure 14

We found that the number of pull requests for each month decreased over time. We have also taken out the month July (last month in the sql dump since there were only 11 total pull requests in the dataset for that month).

## Acknowledgements

- A special word of thanks goes to Professor Devanbu for providing important feedback for the project and helping us improve it.
- I'll also like to thank Casey for helping us through some phases of the project.