2018

Regression Analysis of Open Source Project Impact: Relationships with Activity and Rewards

Vinod Kumar Ahuja
vahuja@unomaha.edu

Regression Analysis of Open Source Project Impact: Relationships with Activity and Rewards

Prepared by:

Vinod Kumar Ahuja

1. Introduction

Engagement with open source projects is becoming an increasingly important part of how people work. In this regard, there is a growing interest in how we can better understand the dynamics within an open source project related to project activity, project contributor rewards, and project impact. In this paper, we summarize our work of exploring the relationships between these items. In particular, we highlight project impact with the following metrics used to define project activity, project contributor rewards, and project impact:

Activity Metrics	
Project Age	Total number of days since the start of a project. (Project Age = project start date – today's date).
Total Watchers	Total number of watchers (subscribers) on a project as of today.
Total Committers	Total number of unique committers on a project as of today.
Total Commits	Total number of commits on a project as of today.
Total Closed Issues	Total number of closed issues on a project as of today.
Pull Request Duration	Median duration in days for all users between the date of their first and last pull request. (Median of all (unique user's pull request duration = first pull request date (for this user) - last pull request date (for this user)).
Pull Request Comment Duration	Median duration in days for all users between the date of their first and last comment on any pull request. (Median of all (unique user's pull request comment duration = first pull request comment date (for this user) - last pull request comment date (for this user)).
Commits Duration	Median duration in days for all users between the date of their first and last commits. (Median of all (unique user's commits duration = first commit date (for this user) - last commit date (for this user))).
Issue Comments Duration	Median duration in days for all users between the date of their first and last comment on any issue. (Median of all (unique user's issue comment duration = first issue comment date (for this user) - last issue comment date (for this user)).
Reward Metrics	
Pull Request Accepted Duration	Median duration in days to accept pull requests. (Pull request accepted duration = pull request open date - pull request last action date (closed or merged, same pull request)).
Pull Request Rejected Duration	Median duration in days to reject pull requests (i.e. closed without merging). (Pull request rejected duration = pull request open date - pull request last action date (closed same issue that does not contain a merged action)).
Average Pull Request Comments	Median number of comments on a project's pull requests.
Impact Metrics	
Upstream Dependencies	Number of other projects the focal project depends on.
Downstream Dependencies	Number of other projects that depend on the focal project.
V-Index	The index of a software that has N number of dependencies in the first order and each dependency in the first order has their own N number of dependencies in second order then the V-Index of that software is N

Table 1. Activity, Reward, and Impact Metrics

2. Methodology

Many metrics are directly available from publically available data sources. However, a new impact metric, the V-index, is calculated as a measure of the impact of open source software as informed by the downstream dependencies of that software. The V-index is calculated in two steps. First, downstream dependencies of a software in question are extracted. This forms the first order dependencies as to how many pieces of software are dependent on the software in question. Second, each extracted downstream dependency is further evaluated and their own downstream dependencies are counted. This forms the second order dependencies. As one piece of software has N number of downstream dependencies in the first order and each of these dependencies in the first order has their own N or more number of downstream dependencies in second order then the V-index of that software is N.

In this paper, we present both (1) the V-index for Mozilla-related projects and (2) the relationship between impact and reward metrics and the V-index. For the calculation of the V-index, data from the Libraries.io database is used. For the calculation of activity and reward variables, data of GHTorrent is used. The data from both the databases are extracted based on following criteria:

1. Projects were selected which started between Jan 1, 2014 to Dec 31, 2015.
2. Rust Programming language is used in all those projects.
3. Data on projects is available in both databases (i.e., Librairies.io and GHTorrent).
4. Outliers removed by dropping all projects with zero dependencies, watchers, or committers.

Based on criteria 1, there were 1,057,169 projects. Based on criteria 2, the list filtered to 10,559 projects. Based on criteria 3, the list filtered to 10,511 projects. Based on criteria 4, the list filtered to 604 projects which are used in this analysis. List of projects and source code is available on GitHub[1].

Next, the V-index was calculated for select projects and a linear regression was performed among different activity, reward, and impact metrics to evaluate which metrics are related and in which direction. For this paper, the V-index is chosen as dependent variable and all other metrics as independent variables.

[1] https://github.com/vinodahujauno/metrics/tree/master/Correlation

3. Findings

a. V-index for Mozilla-related Projects

See Appendix A

b. Correlation Matrix: V-index with Activity and Reward Metrics

	Project Age	Total Watchers	Total Committers	Total Commits	Total Closed Issues	Pull Request Duration	Pull Request Comment Duration	Commits Duration	Issue Comments Duration	Pull Request Accepted Duration	Pull Request Rejected Duration	Avg Pull Request Comments	Upstream Dependencies	Vindex
ProjectAge	1													
TotalWatchers	0.20	1												
TotalCommitters	0.23	0.66	1											
TotalCommits	0.16	0.57	0.80	1										
TotalClosedIssues	0.09	0.59	0.84	0.82	1									
PullRequestDuration	-0.12	-0.11	-0.07	-0.10	-0.09	1								
PullRequestCommentDuration	0.10	-0.02	-0.05	0.00	-0.09	0.21	1							
CommitsDuration	-0.16	-0.21	-0.25	-0.14	-0.17	0.15	0.06	1						
IssueCommentsDuration	-0.06	-0.13	-0.08	-0.12	-0.11	0.20	0.01	0.08	1					
PullRequestAcceptedDuration	-0.08	-0.06	-0.06	-0.06	-0.05	-0.03	-0.01	0.06	0.19	1				
PullRequestRejectedDuration	-0.10	-0.06	-0.06	-0.08	-0.05	0.07	-0.06	0.02	-0.02	0.05	1			
AvgPullRequestComments	0.01	-0.04	-0.06	-0.01	-0.05	-0.03	0.00	0.12	-0.02	-0.02	-0.03	1		
UpstreamDependencies	0.09	0.19	0.16	0.16	0.17	-0.02	0.00	-0.09	0.09	-0.05	-0.03	0.09	1	
Vindex	0.08	0.01	0.02	-0.01	-0.03	-0.04	-0.03	0.00	-0.07	-0.01	0.08	0.03	0.20	1

c. Regression: V-index with Activity and Reward Metrics

V-index ~ Project Age + Total Watchers + Total Committers + Total Commits + Total Closed Issues + Pull Request Duration + Pull Request Comment Duration + Commits Duration + Issue Comments Duration + Pull Request Accepted Duration + Pull Request Rejected Duration + Avg Pull Request Comments + Upstream Dependencies

Residuals:

Min	1Q	Median	3Q	Max
-24.183	-4.181	-2.622	0.449	188.841

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8661822	2.9082563	-0.642	0.5213
Project Age	0.0033336	0.0023827	1.399	0.1623
Total Watchers	-0.0019574	0.0024050	-0.814	0.4160
Total Committers	0.1613861	0.0975651	1.654	0.0986.
Total Commits	-0.0002686	0.0019072	-0.141	0.8881
Total Closed Issues	-0.0173018	0.0091471	-1.891	0.0590.
Pull Request Duration	-0.0064006	0.0107564	-0.595	0.5520
Pull Request Comment Duration	-0.0058760	0.0065690	-0.895	0.3714
Commits Duration	0.0031000	0.0034622	0.895	0.3709
Issue Comments Duration	-0.0204632	0.0093089	-2.198	0.0283*
Pull Request Accepted Duration	0.0068848	0.0309766	0.222	0.8242
Pull Request Rejected Duration	0.0395451	0.0194017	2.038	0.0420*
Avg Pull Request Comments	0.0184534	0.1288367	0.143	0.8862
Upstream Dependencies	0.5121158	0.0938677	5.456	7.1e-08***

Significant Codes: '***' 0.001 // '**' 0.01 // '*' 0.05 // '.' 0.1 // ' ' 1
Residual standard error: 13.6 on 590 degrees of freedom
Multiple R-squared: 0.07185, Adjusted R-squared: 0.0514
F-statistic: 3.514 on 13 and 590 DF, p-value: 2.633e-05

4. Summary

Item	Summary
Activity Metrics	All activity metrics which indicate project size (i.e., age, watchers, committers, commits, closed issues) have a **positive correlation** amongst each other.
	All activity metrics which indicate the duration of activity (i.e., pull request duration, pull request comment duration, commits duration, issue comments duration) have a **negative correlation** amongst each other.
Impact metrics	Upstream and Downstream have **low positive correlation.**
Reward Metrics	All reward metrics (i.e., pull request accepted duration, pull request reject duration, average pull request comments) have **low negative or no correlation** amongst each other.
Activity Metrics vs Reward Metrics	Activity metrics have **no correlation** to reward metrics.
Activity Metrics vs Impact Metrics	Activity metrics have **no correlation** with impact metrics, including the V-index.
Reward Metrics vs Impact Metrics	Reward metrics have **no correlation** with impact metrics, including the V-index.

Table 2 Summary

5. Conclusions

Based on these 604 projects, the V-index, while easily calculable, is not showing any correlation with the chosen activity or reward metrics. Furthermore, the regression equation is not showing significance in its coefficients. Based on our data and these results it can be concluded that impact metrics cannot be increased by improving activity or reward metrics that most of managers use in their daily decisions.

6. Future Work

To increase robustness of our results we are refining our filters. We are extracting the data of activity and reward metrics during the timeframe between the release of two versions of an open source software project. We will explore whether activity and reward data between those two releases impacts the V-index.

7. Acknowledgements

Thanks to Mozilla and the Alfred P. Sloan Foundation for their ongoing support in this work.

Appendix A: V-index for Mozilla-related Projects

See Spreadsheet at: https://goo.gl/9AGwVB