
How Early Participation Determines Long-Term Sustained Activity in GitHub Projects?

.....

In Proceedings of the 31st ACM Joint European Software Engineering
Conference and Symposium on the Foundations of Software Engineering
2023. (ESEC/FSE'23). pp.29–41

OSSプロジェクト

- | 2022年には8,600万のGitHubプロジェクトが作成[7]
 - OSSのメリットが一般的に認識され、プロジェクト数も増加
 - かなりの数のプロジェクトが1年以内に消える[12]
 - 多くの開発者が持続可能性に苦しんでいる[28,80,90]

| OSSの持続可能性

- OSSプロジェクトが複雑なエコシステムを形成
- OSSプロジェクトの失敗が社会へ影響を与える可能性

関連研究

| OSSの持続可能性の決定要因

ー 様々なメトリクスを用いて決定要因を特定

Study	Variables used in the study
Chengalur-Smith and Sidorova [22]	✓: size ✗: age, reliability, niche
Chengalur-Smith et al. [21]	✓: development base size ✗: project age, population niche size
Samoladas et al. [66]	✓: # of developers, project domain
Midha and Palvia [57]	✓: license, # of language translations, responsibility delegation, code complexity, code modularity ✗: user base, developer base
Wang [81]	✓: license, external network size, external network quality, targeted users, user/developer participation ✗: internal network size, service quality
Ghapanchi [39]	✓: defect removal, functionality enhancement ✗: release frequency
Valiev et al. [80]	✓: organization involvement, # of upstream dependencies ✗: # of downstream dependencies, indirect connectivity through transitive dependencies, backporting
Yin et al. [90]	✓: # of developers, # of commits, # of emails, # of files, # of interruptions, % of top contributions ✗: (internal) social network, technical network
Stănculescu et al. [70]	✓: code (e.g., lines of code, # of directories), process (e.g., # of major/minor/new contributors, # of files added/deleted), quality (e.g., tests, code complexity, function size)
Our Study	✓: participant willingness (core/peripheral/non-code), capacity (core/peripheral/non-code), and opportunity

プロジェクト初期で利用できない
変数も多い

- ー OSS参加者の初期の行動と持続可能性の相関を調査[92,93]
 - 初期の参加者要因とプロジェクトの持続可能性との詳細な調査はされていない

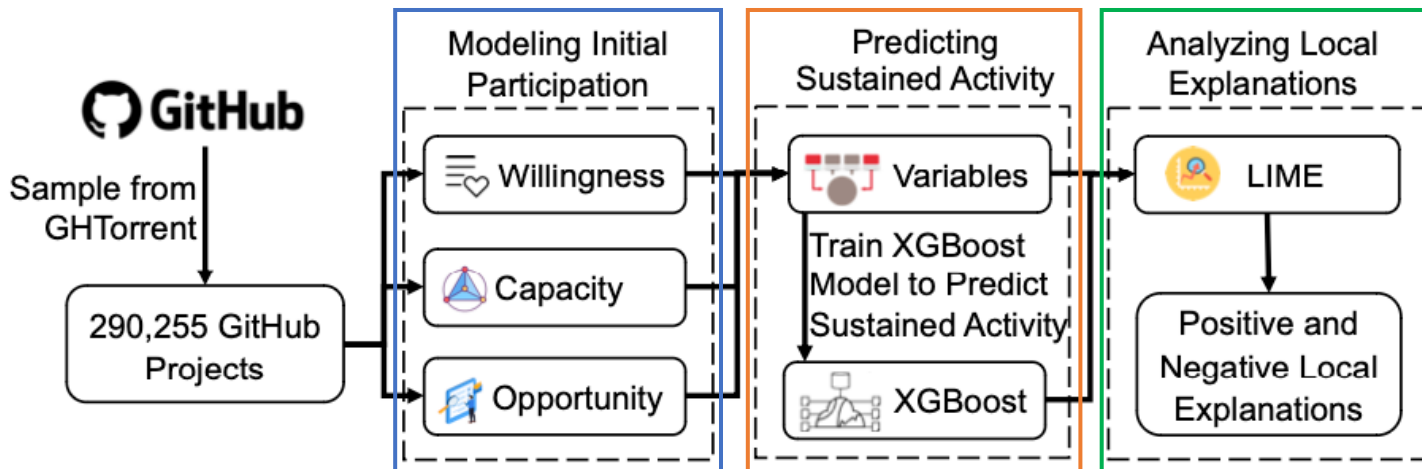
目的 & 概要

目的

初期の参加者要因を用いて将来の持続的な活動を予測する
解釈可能なアプローチを提案

Blumbergモデルを用いて
初期参加を測定する変数を定義

変数と予測の相関を説明



2年間の持続可能性を予測するためのモデルを学習

データセット

| GHTorrent内のプロジェクトをフィルタリング

- フォーク
- 2012年以前 / 2019年3月以降に作成されたプロジェクト
- 長期的な開発を継続するつもりのないプロジェクト

		Ratio of Projects		Number of Projects	
Start	Sum	Two-Year	Still Active	Two-Year	Still Active
2012	17,056	0.46	0.10	7,817	1,779
2013	27,409	0.48	0.11	13,110	2,996
2014	36,226	0.45	0.12	16,359	4,404
2015	46,954	0.43	0.14	20,121	6,458
2016	52,111	0.43	0.17	22,159	9,086
2017	53,863	0.40	0.23	21,278	12,399
2018	49,108	0.39	0.33	19,252	16,235
2019	7,528	0.33	0.33	2,471	2,497
SUM	290,255	0.42	0.19	122,567	55,854

初期参加の測定

| Blumbergモデル

- － 経営科学における行動モデル
- － 意欲, 能力, 機会の要因が個人の行動に影響する

| 開発者の区分

- － 参加者の役割によってプロジェクトへの影響が異なる^[45,47,92]
- － コア開発者 : コミット数が最も多く, 全体の80%以上
- － 周辺開発者 : 少なくとも1つのコミットをした開発者
- － ノーコード開発者 : Issue, PR, コメントのみ

RQ1 | 予測性能

	Variable	Definition	Negative Group	Positive Group	r
Core	#cmt_c	The average number of commits	14.50/18.17	78.00/108.77	0.77↑
	#pr_c	The average number of pull requests	0.00/0.00	0.50/1.14	0.34↑
	#issue_c	The average number of reported issues	4.00/9.04	0.00/0.10	0.81↓
	#iss_comment_c	The average number of issue comments	0.00/0.52	10.00/20.73	0.74↑
	#cmt_comment_c	The average number of commit comments	1.00/2.39	0.00/0.00	0.46↓
	#iss_event_c	The average number of issue events	12.00/33.50	0.00/0.74	0.75↓
	#following_c	The average number of followed developers	47.00/119.59	2.50/5.05	0.75↓
	#star_pro_c	The average number of starred projects	77.00/180.24	4.00/12.99	0.67↓
	#cmt_p	The average number of commits	0.00/0.34	2.90/4.93	0.64↑
	#pr_p	The average number of pull requests	0.00/0.00	0.25/0.56	0.24↑
Peripheral	#issue_p	The average number of reported issues	1.00/2.94	0.00/0.00	0.62↓
	#iss_comment_p	The average number of issue comments	0.00/0.31	0.00/2.11	0.27↑
	#cmt_comment_p	The average number of commit comments	0.00/0.11	0.00/0.03	0.04↓
	#iss_event_p	The average number of issue events	2.00/7.86	0.00/0.04	0.60↓
	#following_p	The average number of followed developers	12.33/41.50	0.00/0.11	0.76↓
	#star_pro_p	The average number of starred projects	24.00/91.02	0.00/0.66	0.75↓
	#issue_n	The average number of reported issues	4.00/9.29	0.00/0.37	0.75↓
	#iss_comment_n	The average number of issue comments	0.00/0.14	1.00/2.38	0.73↑
	#cmt_comment_n	The average number of commit comments	0.00/0.10	0.00/0.03	0.11↓
	#iss_event_n	The average number of issue events	1.20/4.13	0.00/0.39	0.46↓
Non-code	#following_n	The average number of followed developers	1.00/6.12	12.50/49.63	0.26↑
	#star_pro_n	The average number of starred projects	27.90/116.19	1.00/28.90	0.31↑
	#cmt_actday	The number of days with commit activity	6.00/6.00	23.00/27.00	0.87↑
	#cmt_median	The median number of commits per day	0.00/0.13	0.00/0.22	0.03↓
	#cmt_front	The number of commits in the first half of observation period	58.00/89.00	11.00/19.00	0.69↓
	#cmt_end	The number of commits in the second half of observation period	7.00/23.00	9.00/38.00	0.06↓
	cmt_day_std	The standard deviation of commits per day	2.82/3.75	0.85/0.85	0.87↓
	cmt_dev_std	The standard deviation of commits per code contributor	0.00/5.20	17.50/28.58	0.53↑
Core	#cmt_all_c	The average number of commits on all GitHub	162.0/201.0	1541.5/2052.0	0.86↑
	#pr_all_c	The average number of pull requests on all GitHub	44.00/96.77	1.00/9.79	0.74↓
	#issue_all_c	The average number of reported issues on all GitHub	7.00/17.53	50.00/49.88	0.41↑
	#pro_c	The average number of owned projects	20.00/19.00	6.50/6.77	0.83↓
	#pro_oneyear_c	The average number of one-year sustained owned projects	0.00/0.23	3.00/3.53	0.81↑
	#pro_twoyear_c	The average number of owned two-year sustained projects	0.00/0.10	2.00/2.60	0.80↑
	#follower_c	The average number of followers	6.00/7.35	80.00/185.72	0.87↑
	#cmt_all_p	The average number of commits on all GitHub	0.0/17.8	876.5/1462.2	0.75↑
	#pr_all_p	The average number of pull requests on all GitHub	23.00/75.32	0.00/0.09	0.83↓
	#issue_all_p	The average number of reported issues on all GitHub	0.00/0.42	21.60/33.66	0.76↑
Peripheral	#pro_p	The average number of owned projects	15.00/13.99	0.00/0.55	0.78↓
	#pro_oneyear_p	The average number of owned one-year sustained projects	0.00/0.00	1.50/2.17	0.73↑
	#pro_twoyear_p	The average number of owned two-year sustained projects	0.00/0.01	1.00/1.38	0.57↑
	#follower_p	The average number of followers	0.00/6.68	13.00/60.55	0.64↑
	#cmt_all_n	The average number of commits on all GitHub	0.0/200.6	799.7/1230.7	0.55↑
	#pr_all_n	The average number of pull requests on all GitHub	42.00/88.58	0.00/1.27	0.84↓
	#issue_all_n	The average number of reported issues on all GitHub	0.00/5.58	55.00/54.98	0.77↑
	#pro_n	The average number of owned projects	20.00/17.57	0.00/3.76	0.78↓
	#pro_oneyear_n	The average number of owned one-year sustained projects	0.00/0.16	1.00/2.05	0.67↑
	#pro_twoyear_n	The average number of owned two-year sustained projects	0.00/0.86	0.00/0.30	0.25↓
Non-code	#follower_n	The average number of followers	0.00/5.36	55.20/123.72	0.74↑
Oppo.	#iss_open	The number of open issues three months after project creation	3.00/7.00	0.00/2.00	0.42↓
	iss_open_ratio	The ratio of open issues three months after project creation	0.00/0.04	0.63/0.64	0.82↑
	#GFI	The number of good first issues three months after project creation	0.00/2.00	0.00/0.00	0.18↓
	#line_readme	The number of README.md lines	44.00/55.00	23.00/87.00	0.03↓
	#line_contributing	The number of CONTRIBUTING.md lines	0.00/0.00	0.00/40.00	0.23↑
Control Variables	show_comp_c	The ratio of core developers showing affiliated companies/institutions	0.00/0.12	1.00/0.86	0.74↑
	#org_c	The average number of GitHub organizations core developers belong to	1.50/2.30	0.00/0.32	0.52↓
	show_comp_p	The ratio of peripheral developers showing affiliated companies/institutions	0.00/0.00	0.75/0.71	0.74↑
	#org_p	The average number of GitHub organizations peripheral developers belong to	0.00/0.07	0.50/1.22	0.46↑
	show_comp_n	The ratio of non-code contributors showing affiliated companies/institutions	0.00/0.08	0.83/0.69	0.70↑
	#org_n	The average number of GitHub organizations non-code contributors belong to	0.30/1.27	0.00/0.40	0.29↓
	type	The type of project owner account (0: organization, 1: user)	1.00/1.00	0.00/0.00	0.87↓
	#star	The number of stars three months after project creation	5.00/58.00	0.00/0.00	0.77↓
	#fork	The number of forks three months after project creation	1.00/1.00	0.00/9.00	0.12↓
	#member	The number of project members three months after project creation	0.00/0.00	1.00/3.00	0.58↑

r represents the effect size ($r = z/\sqrt{N}$) [37]; ↑, ↓ means the positive group has higher median for this variable and the effect size is small, medium, and large, respectively (↓, ↓↓ vice versa).

開発者の意欲

- 参加者の内発的動機
- 累積量, 安定性, 集中度

参加者の能力

- コード関連の変数
- 一般的なOSS経験, 人気

貢献の機会

- 新規参入する機会の提供
- GFIや充実したREADME

コントロール変数

- プロジェクト固有の特性
- 規模, 人気,

モデルの訓練とLIME

| プロジェクトが t 年以上持続可能かの予測モデル

– XGBoost

- 数値変数とカテゴリ変数の両方をサポート
- ソフトウェア工学の予測タスクで競争力のある性能を示す

| LIME

– 予測結果を説明するためのツール

- 線形/非線形分類器のどちらにも適用可能
- 個々のサンプルの変数寄与を局所的に分析可能

RQ1 | 予測性能

| プロジェクトが t 年間持続可能かを m ヶ月目までのデータを用いて予測

– ハイパーパラメーター：1ヶ月あたりのコミット数 k

| 評価

– AUCによる評価

- 全ての閾値における適合度と予測効果を総合的に測定可能

– 他のモデルとの比較

- ロジスティック回帰, ランダムフォレスト
- 変数の制限

RQ1 | 予測性能

AUC/Precision/Recallによる評価

	<i>m</i>	<i>k</i> = 1	<i>k</i> = 2	<i>k</i> = 6
<i>t</i> = 1	1	0.65 / 0.60 / 0.59	0.66 / 0.58 / 0.39	0.72 / 0.57 / 0.18
	3	0.73 / 0.65 / 0.69	0.74 / 0.63 / 0.56	0.81 / 0.62 / 0.38
	5	0.77 / 0.68 / 0.75	0.80 / 0.67 / 0.65	0.86 / 0.66 / 0.53
<i>t</i> = 2	1	0.66 / 0.59 / 0.41	0.68 / 0.58 / 0.24	0.73 / 0.56 / 0.13
	3	0.72 / 0.62 / 0.52	0.74 / 0.61 / 0.38	0.80 / 0.60 / 0.26
	5	0.76 / 0.65 / 0.59	0.78 / 0.64 / 0.47	0.84 / 0.63 / 0.38

コミット数の
第1・3四分位数, 中央値

ベースラインとの比較

Models	AUC/P/R	Models	AUC/P/R
Logistic Regression	0.67/0.61/0.39	Random Forest	0.65/0.54/0.58
Common Variables	0.66/0.58/0.31	Other Variables	0.71/0.62/0.51
Cumulative Effort	0.59/0.55/0.25	Stability of Effort	0.67/0.58/0.49
Level of Concentration	0.56/0.51/0.15	OSS Experience	0.62/0.57/0.32
Popularity	0.55/0.52/0.17	Opportunity	0.55/0.56/0.09
Control Variables	0.61/0.57/0.30	Our Model	0.72/0.62/0.52

先行研究[70]でも利用されている
変数とそれ以外で比較

RQ2 | 継続的な活動の決定要因

| 参加者の意欲：累積量

- コード貢献者の意欲はプロジェクトの持続可能性に良い影響
 - コミット数, PR数, Issueのコメント数

| 参加者の意欲：安定性

- 安定した定期的なコミットは持続可能性に良い影響
 - コミットが期間の偏りを持つと持続可能性に悪影響

| 参加者の意欲：集中度

- フォロワーの平均数とスターをつけたプロジェクト数が悪影響

RQ2 | 継続的な活動の決定要因

| 参加者の能力：一般的なOSS経験

- － 過去コミットやIssue数は良い影響 / PR数は悪影響
 - ・ 開発者の能力が他のプロジェクトに分散するため
- － 長期的なプロジェクトでの活動経験は良い影響
 - ・ しかしその数が多すぎると役に立たない

| 参加者の能力：人気

- － フォロワー数は良い影響
 - ・ 有用なプロジェクトを開発する能力を表す指標

RQ2 | 継続的な活動の決定要因

| 貢献の機会

- オープンIssueの比率とコントリビューティングファイルの行数は良い影響
 - GFIが多くなるとIssue数も多くなるため悪影響を与える

| コントロール変数

- 組織のプロジェクトは持続可能性が高い
- プロジェクトメンバーの数は良い影響
- スター数やフォーク数は悪影響

持続可能なプロジェクトのキックオフ

| プラクティス

- プロジェクトの明確な目標を設定し，開発者がどのような貢献が可能かを明確にする
 - PRでの議論を少なくできる
- ポジティブなフィードバックを提供
 - 素早くフレンドリーな回答
- 開発者減少の兆候を事前に察知する
 - 貢献の減少と他のプロジェクトへの関心の移行を観察

持続可能なプロジェクトの特定

| 指標

- 貢献者が大規模かつ安定的に貢献している
- コード貢献者は集中的な注意を払っている
- 参加者がコミット，Issue報告，持続可能性の高いプロジェクトの立ち上げ経験を持つ
- 開発者が多くのフォロワーを持っている

補 足 資 料

定義した変数

	Variable	Definition	Negative Group	Positive Group	r	
Willingness of Participants	Core	#cmt_c	The average number of commits	14.50/18.17	78.00/108.77	0.77↑
		#pr_c	The average number of pull requests	0.00/0.00	0.50/1.14	0.34↑
		#issue_c	The average number of reported issues	4.00/9.04	0.00/0.10	0.81↓
		#iss_comment_c	The average number of issue comments	0.00/0.52	10.00/20.73	0.74↑
		#cmt_comment_c	The average number of commit comments	1.00/2.39	0.00/0.00	0.46↓
		#iss_event_c	The average number of issue events	12.00/33.50	0.00/0.74	0.75↓
		#following_c	The average number of followed developers	47.00/119.59	2.50/5.05	0.75↓
		#star_pro_c	The average number of starred projects	77.00/180.24	4.00/12.99	0.67↓
	Peripheral	#cmt_p	The average number of commits	0.00/0.34	2.90/4.93	0.64↑
		#pr_p	The average number of pull requests	0.00/0.00	0.25/0.56	0.24↑
		#issue_p	The average number of reported issues	1.00/2.94	0.00/0.00	0.62↓
		#iss_comment_p	The average number of issue comments	0.00/0.31	0.00/2.11	0.27↑
		#cmt_comment_p	The average number of commit comments	0.00/0.11	0.00/0.03	0.04-
		#iss_event_p	The average number of issue events	2.00/7.86	0.00/0.04	0.60↓
		#following_p	The average number of followed developers	12.33/41.50	0.00/0.11	0.76↓
		#star_pro_p	The average number of starred projects	24.00/91.02	0.00/0.66	0.75↓
	Non-code	#issue_n	The average number of reported issues	4.00/9.29	0.00/0.37	0.75↓
		#iss_comment_n	The average number of issue comments	0.00/0.14	1.00/2.38	0.73↑
		#cmt_comment_n	The average number of commit comments	0.00/0.10	0.00/0.03	0.11↓
		#iss_event_n	The average number of issue events	1.20/4.13	0.00/0.39	0.46↓
		#following_n	The average number of followed developers	1.00/6.12	12.50/49.63	0.26↑
		#star_pro_n	The average number of starred projects	27.90/116.19	1.00/28.90	0.31↓
#cmt_actday		The number of days with commit activity	6.00/6.00	23.00/27.00	0.87↑	
#cmt_median		The median number of commits per day	0.00/0.13	0.00/0.22	0.03-	
Capacity of Participants	Core	#cmt_all_c	The average number of commits on all GitHub	162.0/201.0	1541.5/2052.0	0.86↑
		#pr_all_c	The average number of pull requests on all GitHub	44.00/96.77	1.00/9.79	0.74↓
		#issue_all_c	The average number of reported issues on all GitHub	7.00/17.53	50.00/49.88	0.41↑
		#pro_c	The average number of owned projects	20.00/19.00	6.50/6.77	0.83↓
		#pro_oneyear_c	The average number of one-year sustained owned projects	0.00/0.23	3.00/3.53	0.81↑
		#pro_twoyear_c	The average number of owned two-year sustained projects	0.00/0.10	2.00/2.60	0.80↑
		#follower_c	The average number of followers	6.00/7.35	80.00/185.72	0.87↑
	Peripheral	#cmt_all_p	The average number of commits on all GitHub	0.0/17.8	876.5/1462.2	0.75↑
		#pr_all_p	The average number of pull requests on all GitHub	23.00/75.32	0.00/0.09	0.83↓
		#issue_all_p	The average number of reported issues on all GitHub	0.00/0.42	21.60/33.66	0.76↑
		#pro_p	The average number of owned projects	15.00/13.99	0.00/0.55	0.78↓
		#pro_oneyear_p	The average number of owned one-year sustained projects	0.00/0.00	1.50/2.17	0.73↑
		#pro_twoyear_p	The average number of owned two-year sustained projects	0.00/0.01	1.00/1.38	0.57↑
		#follower_p	The average number of followers	0.00/6.68	13.00/60.55	0.64↑
Non-code	#cmt_all_n	The average number of commits on all GitHub	0.0/200.6	799.7/1230.7	0.55↑	
	#pr_all_n	The average number of pull requests on all GitHub	42.00/88.58	0.00/1.27	0.84↓	
	#issue_all_n	The average number of reported issues on all GitHub	0.00/5.58	55.00/54.98	0.77↑	
	#pro_n	The average number of owned projects	20.00/17.57	0.00/3.76	0.78↓	
	#pro_oneyear_n	The average number of owned one-year sustained projects	0.00/0.16	1.00/2.05	0.67↑	
	#pro_twoyear_n	The average number of owned two-year sustained projects	0.00/0.86	0.00/0.30	0.25↓	
	#follower_n	The average number of followers	0.00/5.36	55.20/123.72	0.74↑	
Oppo.	#iss_open	The number of open issues three months after project creation	3.00/7.00	0.00/2.00	0.42↓	
	iss_open_ratio	The ratio of open issues three months after project creation	0.00/0.04	0.63/0.64	0.82↑	
	#GFI	The number of <i>good first issues</i> three months after project creation	0.00/2.00	0.00/0.00	0.18↓	
	#line_readme	The number of README.md lines	44.00/55.00	23.00/87.00	0.03-	
	#line_contributing	The number of CONTRIBUTING.md lines	0.00/0.00	0.00/40.00	0.23↑	
Control Variables	show_comp	show_comp_c	The ratio of core developers showing affiliated companies/institutions	0.00/0.12	1.00/0.86	0.74↑
		#org_c	The average number of GitHub organizations core developers belong to	1.50/2.30	0.00/0.32	0.52↓
		show_comp_p	The ratio of peripheral developers showing affiliated companies/institutions	0.00/0.00	0.75/0.71	0.74↑
		#org_p	The average number of GitHub organizations peripheral developers belong to	0.00/0.07	0.50/1.22	0.46↑
		show_comp_n	The ratio of non-code contributors showing affiliated companies/institutions	0.00/0.08	0.83/0.69	0.70↑
	#org	#org_n	The average number of GitHub organizations non-code contributors belong to	0.30/1.27	0.00/0.40	0.29↓
		type	The type of project owner account (0: organization, 1: user)	1.00/1.00	0.00/0.00	0.87↓
		#star	The number of stars three months after project creation	5.00/58.00	0.00/0.00	0.77↓
		#fork	The number of forks three months after project creation	1.00/1.00	0.00/9.00	0.12↓
		#member	The number of project members three months after project creation	0.00/0.00	1.00/3.00	0.58↑

r represents the effect size ($r = z/\sqrt{N}$) [37]; ↑, ↓, ↑ means the positive group has higher median for this variable and the effect size is small, medium, and large, respectively (↓, ↓, ↓ vice versa).