OSS プロジェクトにおける Good First Issue PR のマージ予測とリポジトリクラスタリング

S266264

1 背景

オープンソースソフトウェア（OSS）プロジェクトにおいて、新規貢献者のオンボーディングはコミュニティの持続可能性を左右する重要な課題である。多くのプロジェクトは「Good First Issue」（GFI）ラベルを用いて初心者に適したタスクを明示する慣行を採用しており、先行研究では GFI ラベルの利用実態や推奨手法が検討されてきた [1, 2]。

筆者の研究では、37 の GitHub リポジトリにおける 4 年間（2021 年 7 月〜2025 年 6 月）の GFI 慣行を時系列分析しており、GFI 比率の減少傾向やタスクタイプ別マージ率の差異を統計的手法（Mann-Kendall 検定、Mann-Whitney U 検定）で分析している。しかし、これらの統計検定は変数間の**単変量の関連**を個別に評価するものであり、複数の要因が同時にマージ成功に与える影響を捉えることが困難である。

本レポートでは、同じデータセットに対してデータサイエンス的アプローチを適用し、以下の 2 つの分析を行う。

1. **機械学習によるマージ予測**: PR 特性を特徴量として多変量のマージ予測モデルを構築し、単変量検定では見えない非線形関係や交互作用を評価する。
2. **リポジトリクラスタリング**: GFI 慣行の類似性に基づいてリポジトリをクラスタリングし、プロジェクト間の異質性を構造的に分析する。

2 データと方法

2.1 データ概要

データは 2025 年 11 月に GitHub GraphQL API を通じて収集した。分析対象は 37 の GitHub リポジトリ（スター数上位 50 から非ソフトウェアを除外）における 4 年間のデータである。

- **GFI PR**: 1,117 件（30 リポジトリ、全て Bot 除外済み）
- **特徴量**: 追加行数、削除行数、変更ファイル数、説明文長、コミット数、レビューコメント数、タスクタイプ（Bug/Feature/Docs/Other）、分析年度
- **目的変数**: マージ成否（二値分類、マージ率 53.0%）
- **クラスタリング用**: 28 リポジトリの GFI 総数、GFI 比率変化、マージ率、平均レビュー数、PR 数

Table 1: マージ予測モデルの比較（5 分割交差検証）

Model	Accuracy	F1	AUC-ROC
Logistic Regression	0.627	0.656	0.657
Random Forest	0.643	0.674	**0.712**
Gradient Boosting	0.650	0.673	0.702

2.2 解析手法

2.2.1 マージ予測（分類問題）

10 個の特徴量を用いて 3 つの分類モデルを構築した。追加行数・削除行数は対数変換を適用し、全特徴量を標準化した。

- ロジスティック回帰（線形ベースライン）
- ランダムフォレスト（決定木アンサンブル）
- 勾配ブースティング（逐次的アンサンブル）

評価は 5 分割層化交差検証で行い、精度（Accuracy）、F1 スコア、AUC-ROC を算出した。最良モデルの特徴量重要度を分析し、マージ成功に寄与する要因を特定した。

2.2.2 リポジトリクラスタリング

28 の GFI 使用リポジトリについて、5 つの特徴量（GFI 総数、GFI 比率変化率、マージ率、平均レビュー数、PR 数）を Ward 法による階層的クラスタリングで 3 群に分類した。結果を PCA（主成分分析）で 2 次元に射影し可視化した。

3 結果

3.1 マージ予測

表 1に 3 モデルの性能比較を示す。最良モデルはランダムフォレスト（AUC-ROC=0.712）であり、ロジスティック回帰（AUC=0.657）を上回った。ただし、いずれのモデルも AUC-ROC は 0.66〜0.71 の範囲であり、予測性能は**中程度**にとどまった。図 1に ROC 曲線を示す。

勾配ブースティングの特徴量重要度分析（図 2）では、**レビューコメント数**（0.31）が最も高い重要度を示し、次いで**説明文長**（0.14）、**分析年度**（0.12）であった。コード量に関する特徴量（追加行数、削除行数、変更ファイル数）は相対的に低い重要度であった。

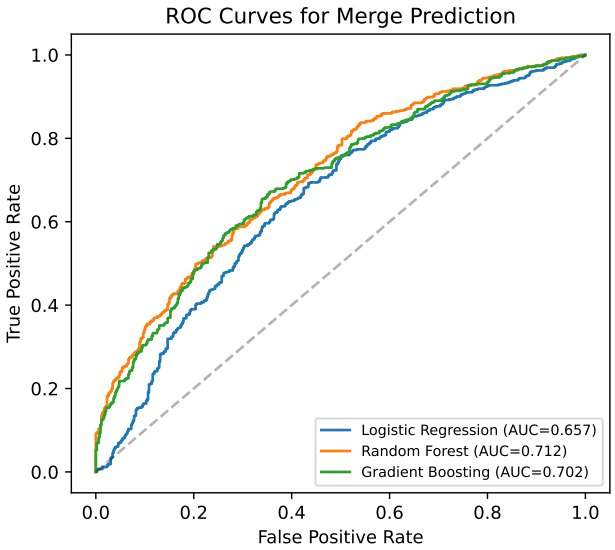

Figure 1: 3 モデルの ROC 曲線。ランダムフォレストが最良の AUC-ROC（0.712）を達成。

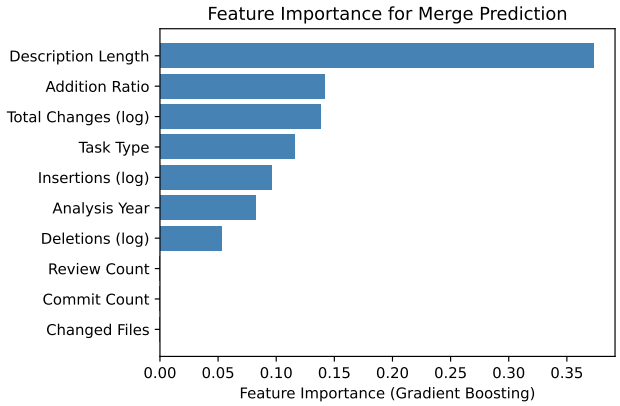

Figure 2: 勾配ブースティングの特徴量重要度。レビューコメント数が突出して高い。

3.2 リポジトリクラスタリング

Ward 法による階層的クラスタリングの結果、28 リポジトリが 3 つのクラスタに分類された（図 3）。

- **Cluster 1**（3 リポジトリ: pytorch, rust, supabase）: GFI 総数が突出して多い（平均 334 件）大規模プロジェクト。GFI 比率は大幅に減少し、マージ率は中程度（53.3%）。
- **Cluster 2**（5 リポジトリ: angular, ant-design, excalidraw 等）: GFI 比率が増加傾向にあるが、マージ率が低い（20.7%）。GFI ラベルの積極的な付与にもかかわらず、PR のマージには至っていない。
- **Cluster 3**（20 リポジトリ）: 最も多数を占め、中程度の GFI 使用量で、マージ率が高い（69.2%）。GFI 慣行が安定的に機能している群である。

3.3 タスクタイプ別マージ率の時系列変化

図 4にタスクタイプ別・年度別のマージ率ヒートマップを示す。Bug-fix タスクは Y2 で 83.5% のピークを

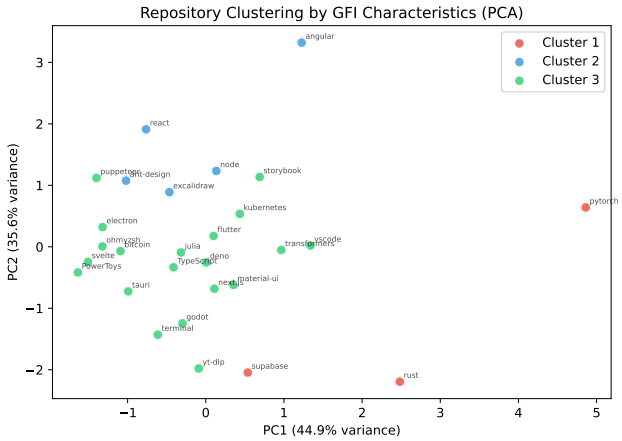

Figure 3: PCA 射影によるリポジトリクラスタリング結果。

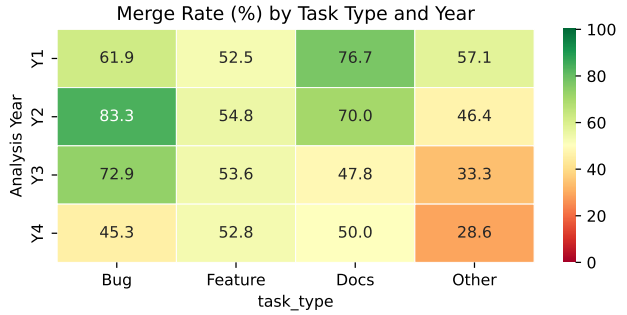

Figure 4: タスクタイプ別・年度別マージ率（%）のヒートマップ。

示した後、Y4 で 45.9% に低下した。Feature タスクは 4 年間を通じて 51〜57% で安定していた。Other カテゴリは 57.1%（Y1）から 28.6%（Y4）まで一貫して減少した。

4 考察

4.1 マージ予測の限界と示唆

機械学習モデルの予測性能が AUC-ROC=0.71 にとどまったことは、PR 提出時点で利用可能な特徴量だけではマージ成否を十分に予測できないことを示す。これは本研究の統計検定（Mann-Whitney U 検定で初期 PR 特性に有意差なし）と整合的であり、**多変量の非線形モデルでも同様の結論**が得られたことで、この知見の頑健性が確認された。

一方、特徴量重要度でレビューコメント数が突出していた点は注目に値する。ただし、レビューコメント数はマージプロセスの**結果**として蓄積される指標であり（branch protection ルールによる機械的相関を含む）、因果的な予測因子とは解釈できない。この交絡を考慮すると、マージ成否は提出時の特性よりも**レビュープロセスにおけるメンタリングの質**に依存する可能性が高く、Cao et al. [3] の知見と一致する。

分析年度が 3 番目に高い重要度を示したことは、マージ率の時系列的減少トレンド（$\tau = -0.35$, $p < 0.001$）がモデルに反映されていることを示す。これは外部環境の変化（品質基準の厳格化等）がマージ成

否に影響している可能性を示唆する。

4.2 クラスタリングの示唆

3 つのクラスタの発見は、GFI 慣行が一様でないことを構造的に裏付けている。特に Cluster 2（GFI 増加・低マージ率）は、GFI ラベルを積極的に付与しているにもかかわらず PR がマージされにくい群であり、ラベリングだけでは新規貢献者の成功に不十分であることを示す。Cluster 3（安定・高マージ率）の多数派としての存在は、適度な GFI 使用と高いマージ率を両立できるプロジェクト運営が可能であることを示唆する。

4.3 データの限界

本分析にはいくつかの限界がある。第一に、GFI ラベルデータは 2025 年 11 月時点のスナップショットであり、ラベル付与・除去の時系列は追跡できていない。第二に、特徴量は PR のメタデータに限定されており、コードの品質や issue の難易度といった質的情報は含まれていない。第三に、クラスタリングの結果は特徴量の選択やクラスタ数に依存しており、異なる設定では異なる結果が得られる可能性がある。

References

[1] H. Tan, et al., “First-timers’ issues: Characterizing and detecting good first issues on GitHub,” *Proc. FSE*, 2020.

[2] T. Turzo, et al., “Evaluating Onboarding Recommendations for Newcomers to Open Source Software Projects,” *IEEE TSE*, 2025.

[3] Y. Cao, et al., “Mentoring Newcomers in Open Source: Insights from a Process Analysis,” *Proc. ICSE*, 2023.

[4] I. Steinmacher, et al., “Social Barriers Faced by Newcomers Placing Their First Contribution in Open Source Software Projects,” *Proc. CSCW*, 2015.