B-028

Web フレームワーク型 OSS を対象にした生存時間の予測と影響要因の分析
Survival Time Prediction and Factor Analysis of Web Framework-related OSS Projects

王 茜 1　邵騰飛 1　岸　知二 1
Xi Wang　Tengfei Shao　Tomoji Kishi

1. はじめに

近年，オープンソースソフトウェア（OSS）は，開発者や企業にとって不可欠な技術資源となっており，広く活用されている．中でも，Web フレームワーク型 OSS は，Web アプリケーション開発の基盤として多くのプロジェクトに採用されており，その選定はシステム全体のアーキテクチャや将来的な保守性に大きく影響を及ぼす．

しかし，OSS の中には短期間でメンテナンスが停止されたり，更新が途絶えたりする例も少なくない．Web フレームワーク型 OSS は更新頻度が高く，技術的競争やコミュニティ依存度も強いため，他の OSS と比べて生存期間が短い傾向にある．そのため，導入時に継続的な利用可能性（持続性）を見極めることが重要である．その生存期間（持続期間）の予測は OSS 選定における重要な判断基準となりうる．

本研究では，Web フレームワーク型 OSS を対象とし，その生存期間に影響を与える要因（特徴量）を分析し，持続可能な OSS 選定を支援することを目的とする．具体的には，GitHub 上の Web フレームワーク関連プロジェクトを収集し，テキストと数値データの両面から特徴量を抽出した上で，生存期間を分類する予測モデルを構築・分析する．

本手法により，OSS の持続性を判断する上での実用的な指標を提示するだけでなく，プロジェクトの初期段階における技術選定や開発体制の構築にも貢献することが期待される．また，データに基づく定量的な分析により，これまで主観的になりがちであった OSS 選定に対して，客観的な意思決定支援の枠組みを提供する．

2. 関連研究

2.1 OSS 生存時間

既存研究では，OSS プロジェクトの生存時間に影響を与える要因の分析が多数行われている．東本ら[1]の研究では，GitHub 上の OSS プロジェクトを対象に，生存時間分析を通じてコミット頻度や開発者数が生存時間に与える影響を明らかにしている．この研究では，プロジェクトの内部属性，すなわち開発活動の頻度や開発メンバー数が，生存時間と有意な関係を持つことが示された．一方で，これらの特徴量を活用して，OSS プロジェクトの生存時間や規模を具体的に予測するモデルの構築までは至っていない．

また，Schweitzer ら[2]は，より広範な OSS プロジェクト群に対する成長パターンの分析を行い，プロジェクト数や規模の拡大がどのような数理的傾向を持つかを明らかにした．この研究では，OSS コミュニティ全体の進化様相をモデル化し，成長速度や構造変化を定量的に評価している．しかしながら，本研究は個々のプロジェクトの持続性ではなく，OSS 全体の集合的挙動に焦点を当てており，特定カテゴリ，たとえば Web フレームワーク型 OSS における生存時間の予測や，それに影響する特徴量の同定を目的としてはいない．

以上の先行研究においては，OSS の生存時間に関する重要な知見が得られているものの，特定のプロジェクト種別に対して予測可能なモデルを構築し，特徴量との関連性を定量的に解釈するアプローチは十分に確立されていない．したがって，本研究では，Web フレームワーク型 OSS に着目し，生存時間予測モデルの構築およびその解釈を通じて，特徴量が生存時間に与える影響を明らかにすることを目的とする．

2.2 順序ロジスティック回帰

既存研究では，順序ロジスティック回帰（Ordered Logistic Regression）は，応答変数が段階的な順序を持つ場合に適した統計的手法として広く用いられている．Hosmer ら[3]の研究では，順序ロジスティック回帰モデルの理論的基盤と応用例について詳細に述べられており，複数のカテゴリにまたがる順序データに対して，しきい値をまたいだオッズ比を推定し，各説明変数が応答に与える影響を定量的に分析できることが示されている．特に，モデルの構築においては，カテゴリ変数のダミー化や多重共線性の処理，モデル適合度の検証手法など，実践的な手法が包括的に提示されている．

一方，自然言語処理の分野においては，テキストから意味的な特徴量を効率的に抽出するための手法として，Sentence-BERT（sBERT）が注目を集めている．特に，Mohammad ら[4]では，学術文献の分類タスクにおいて，パラメータ効率に優れた Transformer ベースの文ベクトル生成モデルとして sBERT を活用し，軽量かつ高精度な分類が可能であることが報告されている．この研究は，従来の BERT ベースモデルに比べて学習資源の削減と性能向上の両立を実現しており，深層学習によるテキスト特徴量抽出の有効性を示している．

したがって，本研究では，OSS プロジェクトの説明文に対して SBERT を用いたベクトル化を行い，それを含む各種特徴量を説明変数とした順序ロジスティック回帰モデルを導入することで，意味的特徴量と数値的特徴量の両面から構成された，より高精度かつ解釈可能な分析手法の実現を目指す．

3. 提案手法

本研究では，Web フレームワークに関連する OSS の生存期間を予測し，その要因を分析する手法を提案する．本手法は，①データ収集と前処理，②BERTopic によるテキスト特徴量抽出，③順序ロジスティック回帰による分類，

1.早稲田大学　University of Waseda

④回帰係数の解釈，の 4 ステップから構成される．（図 1）

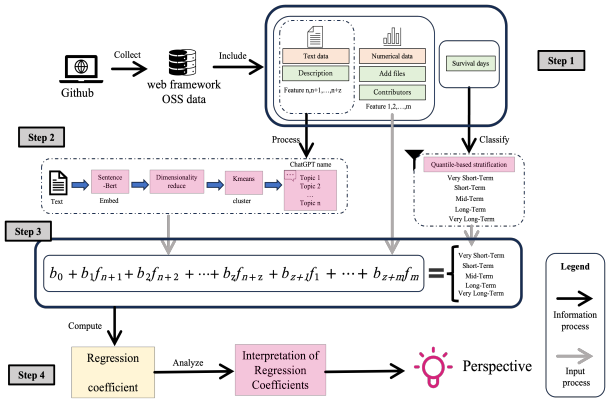

図 1　提案手法

3.1 データ収集と前処理

本研究におけるデータ収集元は GitHub である．GitHub API を利用し，「web-framework」や「framework」「backend」「rest-api」「http-server」「router」などのキーワードに基づき，Web フレームワークに関連すると考えられる OSS プロジェクトをキーワードベースで収集する．

次に，以下の条件を満たすリポジトリにフィルタリングを行う．具体的には，説明文（description）や名前に関連キーワードが含まれていること，topics に該当タグが存在すること，また開発者数が 1 人のみのリポジトリや「browser」「sdk」「template」「demo」などを含む非汎用的なプロジェクトは除外する．

その後，選別されたプロジェクトに対して，説明文（Text data），ファイル追加数，開発者数，ライセンス，最終更新日などの特徴量（Numerical data）を収集する．また，最終的な目的変数として，生存日数（survival_days）を用い，箱ひげ図に基づく五分位数を用いて，Very Short-Term，Short-Term，Mid-Term，Long-Term，Very Long-Term の 5 段階に分類したラベルを作成する．

3.2 Bertopic によるベクトル化

フィルタリング後に残された OSS プロジェクトの説明文を対象とし，まず非英語テキストを英語に翻訳し，前処理を行う．その後，Sentence-BERT [4] を用いて，各プロジェクトの説明文を 384 次元の固定長ベクトルに変換する．

$$v_i = \mathrm{SBERT}(\mathrm{d}_i) \in \mathbb{R}^{384}$$

ここで各説明文dに対して，iは説明文の数を意味する．得られたベクトルは次元削減手法によって，低次元空間へと変換される：

$$v_i' = \mathrm{DR}(v_i) \in \mathbb{R}^K$$

次に，クラスタ数K の決定には Silhouette スコアを用いて適切なクラスタ数を導出し，それに基づいて K-means クラスタリング[5]を実行する．

最後に，各クラスタに属する説明文のうち，距離が近い 5 件を抽出し，それらを ChatGPT に入力することで代表的な概要文を生成し，トピック名を人間可読な形に変換する．

3.3 順序ロジスティックモデル学習

得られた特徴量を用いて，生存期間（目的変数）を 5 段階に分類するための有序ロジスティック回帰モデルを構築する．入力変数としては，①カテゴリ型変数（例：ライセンス）を使用頻度に応じた数値ラベルに変換，②前ステップで得られたテキスト特徴量（クラスタラベル），③数値型特徴量（開発者数，ファイル数など）を用いる．モデルは以下の式で表される：

$$b_0 + b_1 f_{n+1} + b_2 f_{n+2} + \cdots + b_z f_{n+z} + b_{z+1} f_1 + \cdots + b_{z+m} fm = \text{Survival class}$$

ここで，f_i は数値型特徴量，f_{n+z} はクラスタ（トピック）に対応する特徴量を表す．最終的に，有序ロジスティック回帰を用いて以下のように 5 クラスに分類する：

Survival class ∈ {1. Very Short-Term, 2. Short-Term, 3. Mid-Term, 4. Long-Term, 5. Very Long-Term}

このようにして構築されたモデルから回帰係数を解釈することで，OSS プロジェクトの生存期間に与える影響要因を明らかにする知見を得る．

3.4 回帰係数の解釈

構築された有序ロジスティック回帰モデルの学習結果から得られる回帰係数を分析することにより，OSS プロジェクトの生存期間に影響を与える要因を定量的に評価する．

本研究で用いた有序ロジスティック回帰モデル[引用]は，カテゴリ順序を持つ目的変数（Very Short-Term ～ Very Long-Term）に対応しており，各特徴量 x_j に対する回帰係数 β_j の符号および大きさを通じてその影響度を判断する．モデルは以下のように定式化される：

$$\log\left(\frac{\mathrm{P(Y \le k)}}{\mathrm{P(Y > k)}}\right) = \theta_k - \sum_{j=1}^{n} \beta_j x_j \quad (k = 1,\ 2,\ \ldots, K-1)$$

ここで，
Y：生存期間ラベル（5 段階）
θ_k：しきい値パラメータ
β_j：特徴量 x_j に対する回帰係数
K：分類クラス数

各 β_j の符号が正であれば，その特徴量は生存期間が「長くなる」方向に影響を与え，負であれば「短くなる」方向に作用することを意味する．また，係数の絶対値が大きいほど，その変数の影響力は大きいと解釈できる．

以上の 4 ステップのように，本研究ではデータ収集からモデル構築，結果解釈までを一貫して行うことで，OSS プロジェクトの生存期間に影響を与える要因を明らかにすることを目的とした．

4. 実験

本研究では，OSS プロジェクトの生存期間を予測するために，複数の特徴量を収集・整備した．参考にした先行研究で使用されていた特徴量をもとに，本研究ではデータの欠損やノイズを考慮し，最終的に 11 個の特徴量が残った．本研究で用いた特徴量を**表 1** に示す．

次に，生存期間データの分布に対して前処理を行った．生存期間は「最終プッシュ日時 − 作成日時」により算出し，処理前のデータ数は 8396 件であった．外れ値の除去には，四分位範囲（IQR）を用い，「Q1−1.5×IQR〜Q3＋1.5×IQR」の範囲内のデータを抽出対象とした

表 1　生存期間予測に用いた各特徴量の説明

特徴量名	説明
full_name	プロジェクトの名称
stargazers_count	star（スター）数
forks_count	fork（フォーク）数
open_issues_count	オープンな issue の数
add_files	全 commit における 新規追加ファイルの総数
delete_files	全 commit における 削除されたファイルの総数
language	使われたプログラミング言語
contributor_count	コントリビュータ（開発者）数
active_days	commit があった異なる日数
major_contributors	累積 95% の commit を行った最少人数
new_contributors	初回 commit が開始 3 ヶ月以内の開発者数
description	プロジェクトの内容説明
created_at	リポジトリの作成日時
pushed_at	最終プッシュ日時

ただし，Q1−1.5×IQR は負の値となるため，実務的判断として 30 日未満のプロジェクトも除去した．これは，短期間で終了するプロジェクトは開発実態が乏しく，スプリント周期や GitHub の非アクティブ判断基準などからも妥当と考えられる．さらに，description が欠損しているデータも除去し，最終的に 6602 件のデータが得られた．

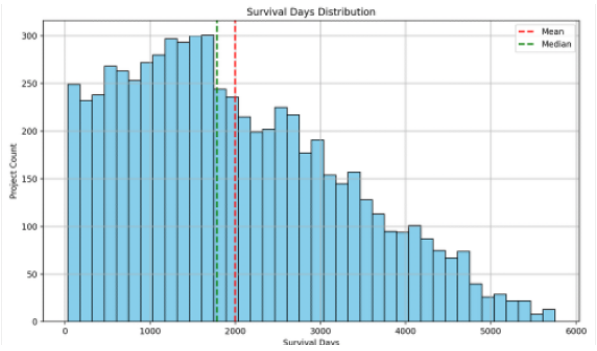

図 2　OSS プロジェクトの生存時間分布(後)

処理前の生存期間分布には、短期間終了プロジェクトが多数含まれていた．図 2 は処理後の分布であり，偏りが軽減され分析に適した形状となっている．得られた生存期間データは，IQR の分位を参考に「短期・やや短期・中期・やや長期・長期」の 5 段階に分類され（表 2），これは後述の分類モデルにおける目的変数（クラスラベル）として使用される．

表 2　生存レベルと日数

生存レベル	生存日数（days）
短期	[30, 961)
やや短期	[961,1783)
中期	[1783,2882)
やや長期	[2882,4320)
長期	[4320,5751)

これらの処理を経て，11 個の妥当な特徴量を有する 6602 件のデータを整備し，生存期間を 5 段階に分類することに成功した．これにより，本研究の目的である OSS プロジェクトの生存期間予測に適した分析基盤が確立された

5. 結果

本節では，クラスタ数の選定およびトピック抽出・モデル学習に関する結果について報告する．

5.1 クラスタ数 K の決定

K-means クラスタリングにおける最適なクラスタ数を決定するため， Silhouette スコアという指標を用いて評価を行った．Silhouette スコアの結果からは，K=11 においてスコアが最大（0.0546）となり，クラスタ間の分離性能が最も高いことが示された．以上の観点から，本研究では K=11 を最適なクラスタ数として採用した．

5.2 トピックの抽出結果

表 3　各トピックと内容概要

トピック名	内容概要
description_topic_0	迅速な機能開発を目的とした シンプルな Web フレームワーク
description_topic_1	Django のユーティリティとコード の簡潔性向上を目的としたヘルパー
description_topic_2	軽量な Flask ベースの RESTful Web アプリケーション
description_topic_3	Rails 風のセマンティックな ルーティングと API 構造
description_topic_4	パフォーマンスとモジュール性に特 化した非同期 マイクロフレームワーク
description_topic_5	Express ベースの Node.js アプリケーション

description_topic_6	aiohttp や Lambda を用いたサーバレス Web アプリケーション
description_topic_7	GraphQL スキーマや API 統合機能
description_topic_8	JSON のシリアライズとバリデーション
description_topic_9	OAuth 認証とセキュアなログインシステム
description_topic_10	FastAPI を基盤としたミドルウェア構成の意見指向型設計

クラスタリングの結果として，11 個のトピックが抽出された．各トピックは，オープンソースソフトウェア（OSS）プロジェクトの説明文から得られた語彙のパターンに基づき，以上の表 3 のような技術的主題に分類された．

これらのトピックにより，プロジェクトの技術的傾向を特徴付けることが可能となり，後の生存期間との関係性分析に寄与する．

5.3 モデル学習結果と特徴量の解釈

構築した分類モデルの性能は，ROC 曲線（図 3）を用いた評価により，平均 AUC 0.5566 を示した．特に，Class 5（生存期間が非常に長い）においては AUC=0.85 と高い識別性能を示した一方，Class 1（生存期間が非常に短い）では AUC=0.28 と識別が困難である傾向が見られた．

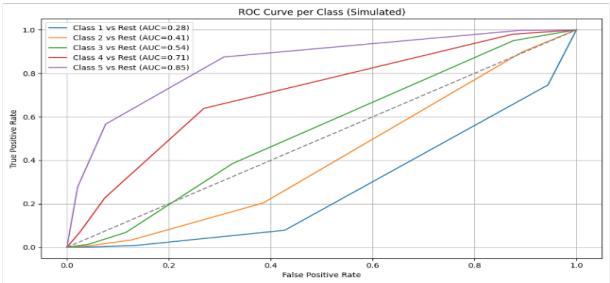

図 3　ROC 曲線図

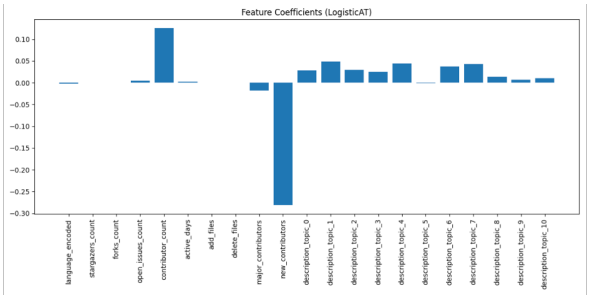

図 4　回帰係数図

回帰係数分析（図４）により，生存期間に対して影響の大きい特徴量が明らかとなった．正の影響を与える特徴量としては「貢献者数」および「description_topic_0～7」が挙げられ，これらは長期にわたり維持されやすいプロジェクトの特徴であると解釈できる．一方，「新規貢献者数」「主要な貢献者数」は負の影響を与える特徴量であり，短期間で終了する傾向にあるプロジェクトを示唆している．

これらの結果は，OSS プロジェクトの持続性を予測する上で有用な特徴量の同定に繋がり，今後のプロジェクト運営や評価指標の設計に貢献するものである．

6. 考察

本研究の分析結果から，OSS プロジェクトの生存期間に影響を与えるいくつかの要因が明らかとなった．

まず，累積貢献者数はプロジェクトの長期的な存続と強く関連しており，継続的な関与が安定した開発体制の構築に寄与していると考えられる．

次に，description テキストから抽出されたトピックのうち，特に description_topic_0～7 に該当するプロジェクトは，比較的長く維持されやすい傾向が見られた．これらは「構造化 API」や「マイクロサービス設計」など，中～大規模開発を志向した内容が中心であり，初期の設計思想が生存期間に影響する可能性を示唆している．

一方，新規貢献者数の多さは必ずしも好影響を与えるとは限らず，急激な増加はプロジェクトの不安定さやメンバー交代の頻発を反映している可能性がある．これは，プロジェクトの継続性に対するリスク要因となり得る．

以上より，開発体制の安定性や初期の方向づけが，OSS プロジェクトの長期的な継続に大きく寄与することが示された．

7. おわりに

本研究では，Web フレームワーク型 OSS プロジェクトを対象に，BERTopic を用いたテキスト特徴量の抽出と有序ロジスティック回帰モデルを組み合わせることで，生存期間の予測およびその要因分析を行った．提案手法により，プロジェクトの説明文に含まれる意味的情報と開発活動に関する数値的特徴量の両面から，生存期間に影響を与える要因を明確化した．

本手法は，OSS プロジェクトの品質・持続性を予測・評価するための基盤を提供し，プロジェクト選定や運用の意思決定に資する実用的な指標の一つとなり得る．今後は，特徴量の拡充や異なる OSS カテゴリへの適用により，より汎用性と実用性に優れた予測モデルの構築を目指す．

参考文献

[1] 東本 知志, 蔵元 宏樹, 斎藤 忍, 飯村 結香子, 近藤 将成, 亀井 靖高, 鵜林 尚靖: 生存時間分析による OSS の活動継続に関する実証評価，第 30 回ソフトウェア工学の基礎ワークショップ（FOSE2023）, pp.55–62 (2023).

[2] Schweitzer, F., et al.: How Do OSS Projects Change in Number and Size? A Large-Scale Analysis to Test a Model of Project Growth, *Advances in Complex Systems*, Vol.17, No.7-8, 1550008 (2014).

[3] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. *John Wiley & Sons*, (2013).

[4] Ahanger, M.M.; Wani, M.A.; Palade, V. sBERT: Parameter-Efficient Transformer-Based Deep Learning Model for Scientific Literature Classification. *Knowledge*, *4*, 397-421(2024).

[5] Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." *Pattern recognition,* 36.2: 451-461. (2003)