УДК 004.412:519.237.5

ПОБУДОВА ТРИФАКТОРНОЇ НЕЛІНІЙНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ JAVA SPRING BOOT ВЕБЗАСТОСУНКІВ

Пухалевич А.В., к.т.н., доц.[1]; Смоквіна В.М.[2]
[1,2] Національний університет кораблебудування імені адмірала Макарова
[1,2] Україна, Миколаїв
[1] andrii.pukhalevych@nuos.edu.ua; [2] 251193@nuos.edu.ua

Анотація. *Вдосконалено нелінійну регресійну модель з трьома предикторами (кількість класів, середня глибина дерева наслідування, середня кількість методів у класі) на основі чотиривимірного перетворення Бокса-Кокса за рахунок знаходження нових значень її параметрів для оцінювання розміру вебзастосунків, що розробляються мовою програмування Java з використанням фреймворку Spring Boot. Це дозволило зменшити значення MMRE та збільшити значення* $PRED(0,25)$ *порівняно з трифакторною моделлю для вебзастосунків на CakePHP та однофакторною моделлю для вебзастосунків на Java та Spring. Для побудови зібрано метрики 82 вебзастосунків з відкритим вихідним кодом. Набір даних було поділено на навчальну вибірку (57 проєктів) і тестову вибірку (25 проєктів). У процесі побудови моделі вилучено 6 спостережень, які виходили за межі інтервалів передбачення. На тестовій вибірці побудована модель має* $R^2 = 0{,}9600$, $MMRE = 0{,}1734$, $PRED(0,25) = 76\%$.

Ключові слова: *оцінювання розміру вебзастосунків; вебзастосунок; Spring Boot; нелінійна регресійна модель.*

Вступна частина. Оцінювання розміру вебзастосунків на Spring Boot є важливою задачею програмної інженерії, оскільки ця інформація використовується на ранній стадії розробки програмного забезпечення для прогнозування трудомісткості за допомогою математичних моделей [1]. Достовірне прогнозування розміру ПЗ дозволяє формувати реалістичні плани, оптимально розподіляти ресурси, визначати строки та бюджет проєкту. Для раннього оцінювання розміру програмного забезпечення використовуються регресійні моделі, у яких залежною змінною є розмір програми в тисячах рядків коду, а незалежними – метрики, що можуть бути отримані з діаграми класів або програмного коду, зокрема загальна кількість класів, середня кількість методів у класі та глибина дерева наслідування [1, 2].

Існуюча модель [2] з такими самими предикторами на основі чотиривимірного перетворення Бокса-Кокса була побудована для вебзастосунків, створених з використанням фреймворку CakePHP. Тобто на даних іншої мови програмування та іншого фреймворку, ніж у даній роботі. Інша існуюча модель [3], призначена для оцінювання розміру вебзастосунків на Java та Spring, має одну незалежну змінну – кількість класів. В ній використовується нормалізуюче перетворення на основі десяткового логарифму, яке нормалізує кожну змінну окремо і не враховує кореляцію між змінними. Крім того, модель [3] охоплює лише невеликі та середні вебзастосунки.

Так як частина функціональності реалізується фреймворком Spring Boot автоматично, то співвідношення між метриками і кількістю рядків коду у вебзастосунках на Spring Boot може відрізнятися від інших Java-застосунків та вебзастосунків, створених з використанням інших фреймворків. Тому постає потреба в підході, який дозволяє врахувати особливості структури коду з впровадженням залежностей в вебзастосунках, розроблених мовою програмування Java із використанням фреймворку Spring Boot, архітектури MVC та характерних шаблонів розробки, притаманних цьому середовищу. Отже, створення нелінійної регресійної моделі для

оцінювання розміру таких застосунків є актуальним та практично значущим завданням.

Метою дослідження є підвищення достовірності оцінювання розміру вебзастосунків, що розробляються мовою програмування Java з використанням фреймворку Spring Boot.

Основна частина. Для побудови нелінійної регресійної моделі було зібрано 82 вебзастосунки з відкритим вихідним кодом з GitHub, що розробляються мовою програмування Java з використанням фреймворку Spring Boot. Набір даних, що містить посилання на репозиторії, хеші та дати проаналізованих комітів і значення метрик усіх 82 проєктів, опубліковано у відкритому доступі на платформі Zenodo [4]. Для оцінювання якості моделі на даних, які не використовувались під час її побудови, набір даних було поділено на навчальну та тестову вибірки у співвідношенні 70% до 30%. У результаті навчальна вибірка містить 57 проєктів, на основі яких будується модель, а тестова – 25 проєктів, які використовуються лише для оцінювання якості моделі.

Коректне застосування лінійного регресійного аналізу можливе лише за виконання низки статистичних припущень, зокрема близького до лінійного зв'язку між залежною та незалежними змінними та нормального розподілу залишків регресії [5]. Дані метрик програмного забезпечення, як правило, цим припущенням не відповідають, оскільки їх розподіл відрізняється від нормального [1, 2]. Тому спочатку необхідно перевірити, чи відповідають емпіричні дані нормальному розподілу. Оскільки модель містить чотири змінні, перевіряється відповідність даних чотиривимірному нормальному розподілу. Для цього використовується тест Мардіа, який ґрунтується на показниках багатовимірної асиметрії та ексцесу [6]. За тестом Мардіа чотиривимірний розподіл даних навчальної вибірки не є нормальним, оскільки статистика асиметрії 313,6648 перевищує критичне значення розподілу χ^2 39,9968. Отже, коректне застосування лінійної регресійної моделі до цих даних неможливе.

Через це було використано метод на основі нормалізуючих перетворень [7], який дозволяє побудувати модель, довірчі інтервали та інтервали передбачення нелінійної регресії для негаусових даних. Як нормалізуюче перетворення використовується чотиривимірне перетворення Бокса-Кокса [2], параметри якого для всіх чотирьох змінних оцінюються одночасно методом максимальної правдоподібності, що дозволяє врахувати кореляцію між змінними під час нормалізації багатовимірних даних [1]. Далі будується лінійна регресійна модель для нормалізованих даних, оцінка коефіцієнтів якої здійснюється методом найменших квадратів. Отримана лінійна модель у нормалізованому просторі трансформується назад у вихідний простір за допомогою зворотного перетворення Бокса-Кокса, утворюючи нелінійну регресійну модель [2]. Спостереження, фактичні значення Y яких виходять за межі інтервалів передбачення нелінійної регресії, вважаються викидами та вилучаються, після чого параметри перетворення Бокса-Кокса оцінюються повторно і модель будується заново. За чотири ітерації було вилучено 6 спостережень, які виходили за межі інтервалів передбачення. П'ять із них мали фактичне значення Y нижче нижньої межі інтервалу передбачення, тобто модель завищувала їх розмір, а одне – вище верхньої межі. На всіх ітераціях чотиривимірний розподіл нормалізованих даних за тестом Мардіа є нормальним, гіпотеза про нормальний розподіл залишків не відхиляється, а значення коефіцієнта збільшення дисперсії не перевищують 1,14, що свідчить про відсутність мультиколінеарності між предикторами. Після підстановки числових значень коефіцієнтів у нелінійну регресійну модель та використання зворотного

перетворення маємо:

$$Y = \left(\left[-5{,}5607 + 1{,}5890 \cdot \frac{X_1^{-0{,}1018} - 1}{-0{,}1018} - 1{,}4188 \cdot \frac{X_2^{-2{,}5650} - 1}{-2{,}5650} + 0{,}6192 \cdot \frac{X_3^{0{,}1549} - 1}{0{,}1549} + \varepsilon \right] \cdot (-0{,}0735) + 1 \right)^{1/(-0{,}0735)}, \quad (1)$$

де Y – розмір вебзастосунку, тисяч рядків коду; X_1 – кількість класів; X_2 – середня глибина дерева наслідування; X_3 – середня кількість методів у класі; ε – випадкова похибка лінійної регресійної моделі для нормалізованих даних.

Прогнозований розмір вебзастосунку $\widehat{Y}$ розраховується за (1) при $\varepsilon = 0$.

Для оцінювання якості побудованої нелінійної регресійної моделі розраховано показники R^2, $MMRE$ (Mean Magnitude of Relative Error), який відображає середню величину відносної помилки між прогнозованими та фактичними значеннями, та $PRED(0,25)$, який показує відсоток прогнозованих значень, величина відносної похибки яких не перевищує заданого порогу. На навчальній вибірці модель має $R^2 = 0{,}9733$, $MMRE = 0{,}1206$, $PRED(0,25) = 94{,}12\%$, а на тестовій вибірці, яка не використовувалась під час побудови моделі, – $R^2 = 0{,}9600$, $MMRE = 0{,}1734$, $PRED(0,25) = 76\%$. Значення $MMRE$ на обох вибірках не перевищує 0,25, а $PRED(0,25)$ не менше 75%, що свідчить про прийнятну якість побудованої моделі [1]. На тій самій тестовій вибірці трифакторна модель [2] зі значеннями параметрів, отриманими для вебзастосунків на CakePHP, має $MMRE = 2{,}0188$ та $PRED(0,25) = 0\%$, а існуюча однофакторна модель для вебзастосунків на Java та Spring [3] – $MMRE = 0{,}6345$ та $PRED(0,25) = 40\%$.

Висновки. У роботі вдосконалено існуючу нелінійну регресійну модель [2] з трьома предикторами на основі чотиривимірного перетворення Бокса-Кокса за рахунок знаходження нових значень її параметрів для оцінювання розміру вебзастосунків, що розробляються мовою програмування Java з використанням фреймворку Spring Boot. Це дозволило зменшити $MMRE$ з 2,0188 до 0,1734 та збільшити $PRED(0,25)$ з 0% до 76% порівняно з моделлю [2]. Використання трьох предикторів та чотиривимірного перетворення Бокса-Кокса дозволило зменшити $MMRE$ з 0,6345 до 0,1734 та збільшити $PRED(0,25)$ з 40% до 76% порівняно з моделлю [3]. Побудовану модель рекомендується використовувати на ранніх етапах розробки Java Spring Boot вебзастосунків в межах значень метрик, на яких її побудовано: кількість класів від 29 до 3528, середня глибина дерева наслідування від 1 до 1,77, середня кількість методів у класі від 1,58 до 11,77.

Використання інструментів ШІ. Під час підготовки тез застосовувався інструмент штучного інтелекту Claude, модель Opus 5.5, для підготовки чорнових варіантів фрагментів тексту. Усі результати перевірено та відредаговано авторами.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Prykhodko, N. V., & Prykhodko, S. B. (2018). The non-linear regression model to estimate the software size of open source Java-based systems. *Radio Electronics, Computer Science, Control*, (3), 158–166. https://doi.org/10.15588/1607-3274-2018-3-17

[2] Prykhodko, S. B., Shutko, I. S., & Prykhodko, A. S. (2021). A nonlinear regression model to estimate the size of web apps created using the CakePHP framework. *Radio Electronics, Computer Science, Control*, (4), 129–139. https://doi.org/10.15588/1607-3274-2021-4-12

[3] Makarova, L., Latanska, L., Pukhalevych, A., Kairov, V., & Dzhurynskyi, M. (2026). Development of a mathematical model for size estimation of Java Spring web applications using regression analysis. *Eastern-European Journal of Enterprise Technologies*, *3*(2 (141)), 67–74. https://doi.org/10.15587/1729-4061.2026.363077

[4] Pukhalevych, A., & Smokvina, V. (2026). *Size and object-oriented metrics of open-source Java*

Spring Boot web applications [Data set]. Zenodo. https://doi.org/10.5281/zenodo.23296810
[5] Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). John Wiley & Sons.
[6] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530. https://doi.org/10.1093/biomet/57.3.519
[7] Prykhodko, S. B., & Prykhodko, N. V. (2021). A modified technique for constructing nonlinear regression models based on the multivariate normalizing transformations. *CEUR Workshop Proceedings*, *3179*, 156–166. https://ceur-ws.org/Vol-3179/Paper_15.pdf

Pukhalevych A.V., Smokvina V.M.

Building a three-factor nonlinear regression model for estimating the size of Java Spring Boot web applications

Abstract. *The nonlinear regression model with three predictors (number of classes, average depth of inheritance tree, average number of methods per class) based on the Box-Cox four-variate transformation is improved by finding new values of its parameters for estimating the size of web applications developed in the Java programming language using the Spring Boot framework. This made it possible to decrease the value of MMRE and increase the value of PRED*(0.25) *compared with the three-factor model for CakePHP web applications and the one-factor model for Java and Spring web applications. Metrics of 82 open-source web applications were collected. The dataset was divided into training (57 projects) and test (25 projects) samples. During model building, 6 observations that fell outside the prediction intervals were removed. On the test sample, the constructed model has* $R^2 = 0.9600$, $MMRE = 0.1734$, $PRED(0.25) = 76\%$.

Keywords: *web application size estimation; web application; Spring Boot; nonlinear regression model.*