УДК 004.412

ТРЬОХФАКТОРНА НЕЛІНІЙНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ DATA SCIENCE ТА MACHINE LEARNING ПРОЄКТІВ, ЯКІ РОЗРОБЛЯЮТЬСЯ ЗА ДОПОМОГОЮ МОВИ ПРОГРАМУВАННЯ JAVA

Орєхов О.С.[1]; Фаріонова Т.А., к.т.н., проф.[2]
[1,2] Національний університет кораблебудування імені адмірала Макарова
[1,2] Україна, Миколаїв
[1]ORCID 0000-0002-0001-0140; [2]ORCID 0000-0003-3384-4712

Анотація. *Робота присвячена побудові трьохфакторної нелінійної регресійної моделі для оцінювання кількості рядків коду проєктів Data Science та Machine Learning, що створюються за допомогою мови програмування Java в залежності від метрик кількості класів, кількості видимих методів та загальна кількість видимих полів класів із використанням перетворення Бокса-Кокса.*

Ключові слова: *нелінійна регресійна модель; нормалізуюче перетворення; data science; java; оцінювання розміру програмних проєктів.*

Вступна частина. У сучасному світі підприємства та організації все активніше застосовують у своїй діяльності проєкти у сфері Data Science та Machine Learning. Ті підприємства, які не використовують переваги даного напрямку галузі інформаційних технологій, ризикують втратити свою конкурентоспроможність на ринку. Це підштовхує підприємства інвестувати час та фінанси в розробку проєктів Data Science та Machine Learning. Створення таких проєктів має більше факторів невизначеності ніж розробка проєктів в інших напрямках, таких як прикладне ПЗ чи розробка веб-застосунків.

В Data Science та Machine Learning використовується широкий набір мов програмування та технологій, в тому числі мова програмування JAVA, яка є однією з найбільш популярних та поширених у світі. Для JAVA існує широкий набір інструментарій для розробки проєктів напрямку Data Science та Machine Learning [1]. Зокрема JAVA використовується для розробки великих enterprise проєктів. А отже, перед підприємствами які, мають широке використання JAVA у своїй інфраструктурі, постають задачі інтеграції відповідних Data Science та Machine Learning проєктів у свої системи. Оцінка прогнозування трудомісткості та витрат розробки таких проєктів є важливою задачею на ранніх етапах проєктування. Достовірні оцінки трудомісткості дозволяють провести раціональний розподіл ресурсів підприємства та оцінити час розробки програмного забезпечення. Для прогнозування трудомісткості розробки використовуються такі математичні моделі, як COCOMO, COCOMO II, ISBSG, SEER-SEM в яких застосовується параметр розміру, а саме кількість рядків програмного коду [2].

Авторами проведено аналіз існуючих моделей оцінки розміру JAVA застосунків на основі метрик програмного коду, а саме чотирьохфакторної нелінійної регресійної моделі для оцінювання розміру JAVA-застосунків з відкритим кодом [3], нелінійної регресійної моделі для оцінювання тривалості розробки програмного забезпечення на JAVA для ПК та платформи MIDRANGE [4,5], нелінійної регресійна модель для оцінювання розміру програмного забезпечення систем з відкритим кодом на JAVA [6]. Серед існуючих моделей для JAVA було виявлено, що якість та точність оцінювання параметру рядків коду залежить від категорії програмного проєкту, фреймворків та бібліотек, метрик які використовуються для оцінювання та деякі моделі не мають можливості оцінити кількість рядків коду та використовуються для оцінки трудомісткості розробки на основі інших параметрів.

Проведений аналіз літературних джерел показав відсутність параметричних моделей для оцінювання розміру Data Science та Machine Learning проєктів, які розробляються за допомогою мови програмування JAVA. Тому розробка нелінійної регресійної моделі для таких проєктів є актуальною.

Метою дослідження є побудова нелінійної трьохфакторної регресійної математичної моделі для оцінювання розміру JAVA-проєктів Data Science та Machine Learning на основі метрик, які можна отримати на етапі проєктування, а саме - кількість класів, кількість видимих методів та загальна кількість видимих полів класів, що дозволить підвищити достовірність оцінювання розміру відповідних проєктів.

Основна частина. Розглянемо побудову нелінійної регресійної моделі для оцінювання кількості рядків коду проєктів Data Science та Machine Learning. Для цього було зібрано дані з 69 проєктів, які розміщені на веб-сервісі Github (https://github.com). Для цих даних отримані метрики аналізу коду за допомогою інструменту CK (https://github.com/mauricioaniche/ck). Для побудови моделі були обрані наступні метрики kClass (кіло-class), kTVMQ (кіло - total visible methods quantity), kPFQ (кіло - public fields quantity), які відповідають факторам X_1, X_2, X_3. Фактори були перевірені на наявність мультиколінеарності за коефіцієнтом впливу дисперсії (VIFs) та мають такі значення - 3.549, 0.098, 4.548, що свідчить про відсутність мультиколінеарності. Перевірка чотирьохвимірних даних на відповідність нормальному закону розподілу багатовимірних даних за критерієм Мардія, показала що вони не розподілені за нормальним законом.

На основі факторів X_1, X_2, X_3 було побудовано лінійну регресійну модель:

$$y = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + \varepsilon, \quad (1)$$

де $\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3$ - параметри оцінки лінійної регресії, отримані за методом найменших квадратів, ε – випадкова величина, яка розподілена за нормальним законом розподілу.

Нульова гіпотеза про нормальність розподілу моделі (1) була перевірена за критерієм Пірсона. Ця гіпотеза була відкинута для рівня значимості a=0,005, для якого значення $\chi^2 = 14.952$, що перевищує значення $\chi^2_{\text{крит}} = 12.838$. Отже, випадкова величина ε не розподілена за нормальним законом, як наслідок це вказує на відсутність теоретичного обґрунтування для використання лінійної регресійної моделі. Отже, постає необхідність у побудові нелінійної регресійної моделі.

Для багатовимірних даних було обране нормалізуюче перетворення Бокса-Кокса та побудовано еліпсоїд прогнозування для нормалізованих даних для рівня значущості a=0.005. Параметри для багатовимірних даних були оцінені за допомогою логарифмічної функції максимальної правдоподібності перетворення Бокса-Кокса для даних із високими коефіцієнтами кореляції. Коефіцієнти нормалізуючого перетвоерення Бокса-Кокса прийняли наступні значення для залежної величини Y та факторів X_1, X_2, X_3 – $\lambda_y = 0.087$, $\lambda_1 = 0.131$, $\lambda_2 = 0.114$, $\lambda_3 = 0.0106$. Під час перевірки даних на викиди, було відкинуто значення з 2-х програмних проєктів.

На основі нормалізованих даних побудована лінійна регресійна модель

$$Z_y = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \varepsilon, \quad (2)$$

де $\hat{b}_0$= -0.555, $\hat{b}_1$= 4.15, $\hat{b}_2$= 10.499, $\hat{b}_3$=6.902 - параметри оцінки лінійної регресії, отримані за методом найменших квадратів.

Далі, на основі моделі лінійної регресії для нормалізованих даних, за допомогою зворотного перетворення було побудовано нелінійну регресійну модель:

$$y = \lambda_y\left(\hat{b}_0 + \frac{\hat{b}_1}{\lambda_1}(X_1^{\lambda_1} - 1) + \frac{\hat{b}_2}{\lambda_2}(X_2^{\lambda_2} - 1) + \frac{\hat{b}_3}{\lambda_3}(X_3^{\lambda_3} - 1) + \varepsilon\right)^{\frac{1}{\lambda_y}}, \qquad (3)$$

де $\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3$ - параметри оцінки лінійної регресії, отримані за методом найменших квадратів, ε – випадкова величина, яка розподілена за нормальним законом розподілу.

Для моделі (3), що була побудована за 67 проектами на основі нормалізуючого перетворення Бокса-Кокса, значення R^2, MMRE і PRED(0,25), дорівнюють відповідно 0.908, 0.192 і 0.681. Слід зазначити що за параметром PRED(0,25) результат моделі перевищує значення 0.25, це свідчить про необхідність проведення досліджень з удосконалення трьохфакторної нелінійної регресійної моделі для оцінювання розміру Data Science та Machine Learning проєктів, які розробляються за допомогою мови програмування JAVA.

Висновки. 1. Побудовано нелінійну трьохфакторну регресійну математичну модель для оцінювання розміру JAVA-проєктів Data Science та Machine Learning на основі метрик програмного коду kClass, kTVMQ, kPFQ. 2. Перспектива подальших досліджень пов'язана з удосконаленням трьохфакторної нелінійної регресійної моделі для оцінювання розміру проєктів в області Data Science та Machine Learning, які розробляються за допомогою мови програмування JAVA, за рахунок використання нормалізуючих перетворень Джонсона.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Haq, H. B., Kayani, H., Toor, S. K., Zafar, S. & Khalid, I. (2020). The Popular Tools Of Data Sciences: Benefits, Challenges and Applications. *International Journal of Computer Science and Network Security, 20*(5)*,* 64-75.

[2] Munialo, S. W. (2016). A Review of Agile Software Effort Estimation Methods. *International Journal of Computer Applications Technology and Research. Association of Technology and Science*, *9*, 612-618. http://dx.doi.org/10.7753/IJCATR0509.1009

[3] Приходько, С. Б., Приходько, Н. В. & Смикодуб, Т. Г. (2020). Чотирьохфакторна нелінійна регресійна модель для оцінювання розміру JAVA-застосунків з відкритим кодом. *Інформатика, обчислювальна техніка та автоматизація*. *31*(70), ч. 1, № 2, 158-162.

[4] Prykhodko, S. B., Pukhalevych, A. V., Prykhodko, K. S. & Makarova, L. M. (2022). Nonlinear regression models for estimating the duration of software development in JAVA for PC based on the 2021 ISBSG data. *Радіоелектроніка, інформатика, управління, 3*, 144-153.

[5] Приходько, С. Б. & Фаріонова, Т. А. (2021). Нелінійні регресійні рівняння для оцінювання тривалості розробки JAVA-застосунків для платформи MIDRANGE. *ITMAS - 2021. Інформаційні Технології: Моделі, Алгоритми, Системи*, *2*, 20-23.

[6] 6. Prykhodko, N. V. & Prykhodko, S. B. (2018). The non-linear regression model to estimate the software size of open source JAVA-based systems. *Радіоелектроніка, інформатика, управління, 3*, 158-166.

Oriekhov O. S., Farionova T.A.

Three-factor nonlinear regression model for estimating the size of Data Science and Machine Learning projects created using the JAVA programming language

Abstract. *This work is dedicated to constructing a three-factor nonlinear regression model for estimating the number of lines of code of Data Science and Machine Learning projects created using the Java programming language depending on the metrics of the number of classes, the number of visible methods, and the total number of visible class fields using the Box-Cox transformation.*

Keywords: *nonlinear regression model; normalizing transformation; data science; Java; software development cost estimation.*