УДК 512.44
DOI 10.24144/2616-7700.2022.1(40).168-174

Н. Е. Кондрук

ДВНЗ «Ужгородський національний університет»,
доцент кафедри кібернетики і прикладної математики,
кандидат технічних наук
natalia.kondruk@uzhnu.edu.ua
ORCID: https://orcid.org/0000-0002-9277-5131

МОДЕЛІ БАГАТОФАКТОРНОГО ПРОГНОЗУВАННЯ

Дане дослідження є розвитком напрямку прикладного аналізу даних. Він відіграє важливу роль у виявленні значущої інформації в наборах даних, яка допомагає приймати обґрунтовані рішення в різних сферах людської діяльності. Наведено інформаційні технології багатофакторного прогнозування, які базуються на моделях MLR та DR і є частиною класичного машинного навчання. Розроблена інформаційно-аналітична система на мові програмування Python та бібліотеки scikit-learn, що реалізує описаний підхід. В якості апробаційної моделі обрана актуальна задача прогнозування ВВП України за показниками: індекс інфляції, чисельність населення, офіційний курс долара, рівень безробіття у відсотках та міграційний приріст. Навчальна вибірка містила 16 спостережень. В ході експериментального дослідження кращою виявилось модель дерева регресії із показником коефіцієнту детермінації 99% та середньої абсолютної відсоткової похибки 6%. Дані індекси якості моделі вказують на її високу точність. Перспективні дослідження полягають у розвитку підходу прикладного аналізу даних для розв'язання різних видів прикладних задач.

Ключові слова: багатофакторне прогнозування, дерева регресії, багатофакторний лінійний аналіз, прогнозування ВВП.

1. Вступ. В наш час все важливішим стає аналіз даних різної природи, тобто процес застосування статистичних та логічних методів для опису, візуалізації, скорочення, перегляду, узагальнення та оцінки даних. Він відіграє важливу роль у виявленні значущої інформації в наборах даних, яка допомагає приймати обґрунтовані рішення.

Ціль будь-якого аналізу даних полягає в тому, щоб отримати з необробленої інформації точну оцінку. Однією із найбільш важливих та поширеніших проблем є встановлення наявності статистичного зв'язку між змінною відгуком (Y) та незалежними (пояснювальними) змінними предикторами (X_i). Дану проблему можна вирішити провівши регресійний аналіз. Вид регресійної моделі залежить від типу розподілу Y: якщо він неперервний та приблизно нормальний, використовується модель лінійної регресії; якщо дихотомічний, то – логістичну регресію; якщо пуасонівський чи поліноміальний необхідним є логарифмічно-лінійний аналіз [1, 2]. Обрана модель «намагається» передбачити результат (Y) на основі значень набору змінних-предикторів (X_i).

Багатофакторна лінійна регресія — це процедура, яка оцінює коефіцієнти лінійного рівняння за участю кількох предикторів, які найкраще передбачають значення залежної кількісної змінної [3].

Іншим підходом до проведення багатофакторного аналізу даних є дерева прийняття рішень. Дерева класифікації та регресії — це методи класичного машинного навчання для побудови моделей прогнозування на основі навчальних даних [4]. Реалізуються такі моделі шляхом рекурсивного поділу простору

даних та підбору методу прогнозування у кожній підмножині. В результаті, розбиття може бути представлено графічно, як дерево рішень. Дерева регресії призначені для визначення значення цільової змінної, яка є неперервною та числовою [5].

Задача розрахунку прогнозного значення ВВП є багатофакторною, бо щонайменше має включати особисті витрати населення на кінцеве споживання товарів та послуг, державні витрати на купівлю товарів та послуг, валові інвестиції та чистий експорт [6]. В даному дослідженні ставиться задача підбору моделі прогнозування величини ВВП України на основі предикторів, які опосередковано впливають на його формування.

2. Багатофакторний (множинний) лінійний аналіз [7]. (Multivariate Linear Regression, MLR). Мета регресії – передбачити Y на основі X або описати, як Y залежить від X. X_i (X_1, X_2, ..., X_k) визначається як "предикторні "пояснювальні"або "незалежні"змінні, тоді як Y називають "залежною змінною "реакцією"або "результатом".

Математичне рівняння моделі:

$$y_i = b_0 + \sum_{j=1}^{n} b_j x_{ij} + e_i,$$

y_i – фактичні значення залежної змінної;

b_0 – коефіцієнт перетину;

b_j – коефіцієнт нахилу (середнє збільшення результату на одиницю збільшення предиктора);

e_i – похибки моделі.

Коефіцієнт детермінації це доля дисперсії прогнозованої змінної, яка пояснюється розглядуваною моделлю:

$$R^2 = \frac{\sum\limits_{i=1}^{n} e_i^2}{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2}.$$

Він може приймати значення від нуля до одиниці. Чим ближче значення коефіцієнта до 1, тим сильнішою є залежність. При оцінці регресійних моделей це інтерпретується як відповідність моделі вхідним даним. В загальному, R^2 має бути не меншим ніж 0,5. Моделі із коефіцієнтом детермінації більшим за 80% можна визначити як достатньо добре підігнані. Основним недоліком використання R^2 є те, що його значення не зменшується із додаванням в модель регресорів навіть, якщо вони ніякого відношення до залежної змінної не мають. Для зняття цього недоліку користуються скоригованим коефіцієнтом детермінації:

$$R^2_{adj} = 1 - \left(1 - R^2\right) \frac{n-1}{n-k},$$

де k – кількість параметрів, n – кількість спостережень.

На відміну від R^2 скоригований коефіцієнт може бути і від'ємним.

3. Дерева регресії [8]. (Regression Tree, RT). Дерева прийняття рішень це методи, які описують правила класифікації в ієрархічній структурі, що складається з двох типів елементів – вузлів та листів. У вузлах розміщені правила

класифікації та виконується перевірка відповідності спостережень цьому правилу за деяким атрибутом навчальної вибірки. В найпростішому випадку в результаті перевірки спостереження що знаходяться у вузлі розбиваються на дві підмножини: ті що задовольняють правила, та ті що ні. Далі кожній підмножині ставиться у відповідність інше правило і процедура рекурсивно повторюється, доки не досягається умова зупинки алгоритма (коли лист містить одне спостереження, або задовольняється умова обмеження на допустиму глибину дерева). В результаті, в останньому вузлі перевірка і розбиття не проводиться і він стає листом. Для дерева регресії кожен лист відповідає деякому значенню прогнозованої змінної. Для оцінки прогностичної сили дерева регресії використовується також коефіцієнт детермінації.

При визначенні точки розбиття у певному вузлі неперервного атрибуту регресійне дерево обирає ту, яка мінімізує середньоквадратичну похибку:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{\text{прог}})^2,$$

де $y_i^{\text{прог}}$ – прогнозне значення залежної змінної.

4. Індекс оцінки якості моделі. Для оцінки якості моделі додатково будемо користуватись середньою абсолютною похибкою в процентах MAPE:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i^{\text{прог}}}{y_i}\right|.$$

5. Експерименти.

1. Постановка задачі. Для проведення практичних експериментів обрано задачу прогнозування ВВП України. Відомо, що існує ряд параметрів, які безпосередньо є складовими даного показника: споживчі витрати, валове нагромадження, експорт та імпорт товарів та послуг [6]. Тому для прогнозування обрано інші показники, які можуть опосередковано впливати на нього: індекс інфляції, чисельність населення, офіційний курс долара, рівень безробіття у відсотках та міграційний приріст. Створено навчальну вибірку із 2005 по 2020 рік на основі відкритих даних Держкомстату та Мінфіну України [6, 9]. Для розв'язання поставленої задачі була створена інформаційно-аналітична система на мові програмування Python.

2. Модель MLR. Для початку оцінимо силу кореляційного зв'язку між ознаками за тепловою картою (рис. 1).

Найбільше із «ВВП» корелюють «Чисельність населення» та «Курс долара», але між собою ці ознаки також мають високий коефіцієнт кореляції, хоча змістовно вони не пов'язані. Це буває у тих випадках, коли динаміка зміни процесів схожа. Всі інші ознаки можемо вважати достатньо незалежними. Враховуючи теплову карту, пропонується навчати модель поступово додаючи предикторні змінні у наступній послідовності: X_1 – «Чисельність населення», X_2 – «Рівень безробіття», X_3 – «Індекс інфляції», X_4 – «Міграційний приріст» і на кінець X_5 – «Курс долара». Залежна змінна Y – «ВВП». При цьому будемо обчислювати два показника якості моделі R^2_{adj} та R^2.

Із наведеної таблиці видно, що скоригований параметр детермінації досягає найбільшого значення при чотирьох предикторних змінних, тому змінна X_5 не

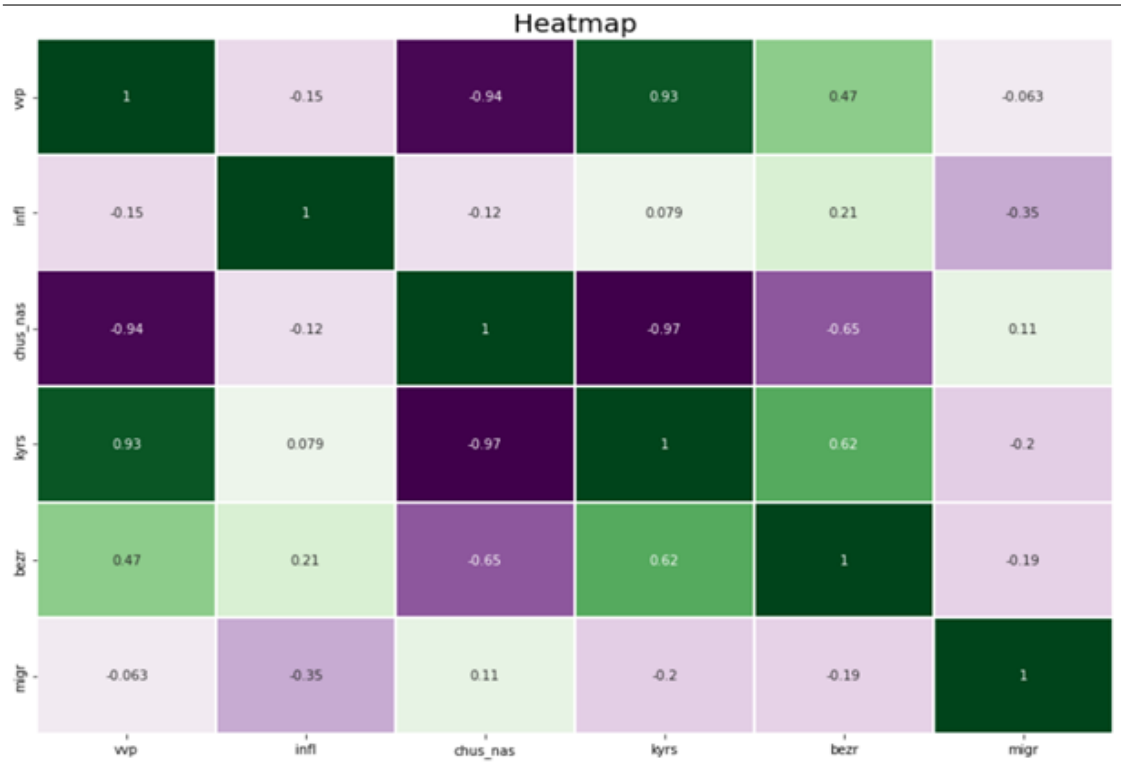

Рис. 1. Теплова карта кореляційного зв'язку між ознаками.

Таблиця 1.

Значення коефіцієнтів детермінації для MLR моделі

Предикторні змінні моделі	X_1	X_1, X_2	X_1, X_2, X_3	X_1, X_2, X_3, X_4	X_1, X_2, X_3, X_4, X_5
R^2	0.876754	0.908755	0.965630	0.969940	0.970321
R^2_{adj}	0.876754	0.902238	0.960342	0.962425	0.959529

покращує MLR модель і виявилась зайвою. Коефіцієнт R^2 при цьому вказує, що варіація предикторних змінних X_1, X_2, X_3, X_4 на 97% призводить до зміни значення Y, тобто модель дуже добре підігнана.

3. Модель RT. При побудові дерев регресії для зупинки алгоритму будемо враховувати параметр максимальної глибини дерева *max_depth*.

Таблиця 2.

Значення коефіцієнту детермінації для DT моделі

Параметр *max_depth*	2	3	4
R^2	0.95695	0.99499	0.99929

Очевидно (табл. 2), що найоптимальніший параметр глибини дерева регресії рівний трьом шарам (рис.2).

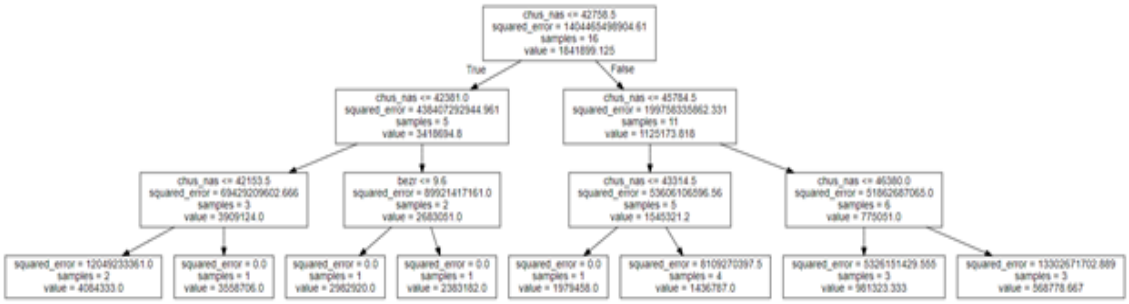

Рис. 2. Дерево регресії глибини 3.

Для побудови дерева використано лише два параметри: «Чисельність населення» та «Рівень безробіття».

4. Оцінка якості моделей. Обчислимо індекси якості моделей та візуалізуємо отримані прогнози.

Таблиця 3.

Значення індексів якості моделей

Модель	MLR	RT
MSE	42217110036	7026376106
MAPE	10,7%	6%

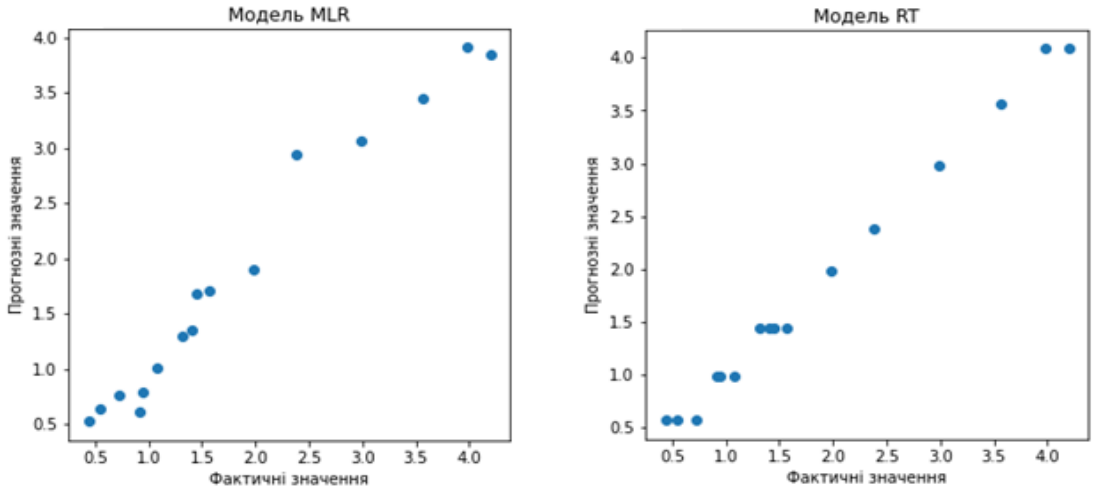

Рис. 3. Візуалізація результатів моделей.

Очевидно, що модель RT краще підігнана до вхідних даних, що підтверджено індексами якості моделі (табл. 3) та візуалізацією результатів (рис. 3).

6. Висновки та перспективи подальших досліджень. Дане дослідження є розвитком напрямку прикладного аналізу даних [10–13].

Наведено інформаційні технології багатофакторного прогнозування, які базуються на моделях MLR та DR і є частиною класичного машинного навчання. Розроблена інформаційно-аналітична система на мові програмування Python

та бібліотеки scikit-learn, що реалізує описаний підхід. В якості апробаційної моделі обрана актуальна задача прогнозування ВВП України за показниками: індекс інфляції, чисельність населення, офіційний курс долара, рівень безробіття у відсотках та міграційний приріст. Навчальна вибірка містила 16 спостережень. В ході експериментального дослідження кращою виявилось модель дерева регресії (RT) із показником коефіцієнту детермінації 99% та середньої абсолютної відсоткової похибки 6%. Дані індекси якості моделі вказують на її високу точність.

Перспективні дослідження полягають у розвитку підходу прикладного аналізу даних для розв'язання різних видів прикладних задач.

Список використаної літератури

1. Gogtay N. J., Deshpande S. P., Thatte U. M. Principles of regression analysis. Journal of the Association of Physicians of India. 2017. Vol. 65(48). P. 48–52.
2. Daoud, Jamal I. Multicollinearity and regression analysis. Journal of Physics: Conference Series. IOP Publishing. 2017. Vol. 949, No. 1. P. 1–6. DOI: https://doi.org/ 10.1088/1742-6596/949/1/012009
3. Shrestha, Noora. Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics. 2020. Vol 8(2). P. 39–42. DOI: https://doi.org/ 10.12691/ajams-8-2-1
4. Torgo L. Regression Trees. In: Sammut C., Webb G. Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. 2017. DOI: https://doi.org/10.1007/978-1-4899-7687-1_717
5. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. Routledge. 2017. DOI: https://doi.org/10.1201/9781315139470
6. Мінфін України. URL: https://minfin.com.ua/ua/
7. Ayyadevara V., Kishore. Linear regression. Pro Machine Learning Algorithms. Apress, Berkeley, CA. 2018. Pp. 17–47. DOI: https://doi.org/10.1016/j.eswa.2017.04.003
8. Song Y. Y., Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015. Vol. 27(2). Pp. 130–135. DOI: https://doi.org/10.11919/j.issn.1002-0829.215044
9. Держкомстат України. URL: http://www.ukrstat.gov.ua/
10. Кондрук Н. Е. Використання мір подібності в методах класифікаціїі. *Науковий вісник Ужгородського університету : серія: Математика і інформатика.* 2021. Вип. 1(38). С. 143–148. DOI: https://doi.org/10.24144/2616-7700.2021.38(1).143-148
11. Kondruk N. E. Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control.* 2018. 3(46). P. 98–105. DOI: https://doi.org/10.15588/1607-3274-2018-3-11
12. Kondruk N. E., Malyar M. M. Analysis of Cluster Structures by Different Similarity Measures. Cybernetics and Systems Analysis, 2021. Vol. 57. Pp. 436–441. DOI: https://doi.org/10.1007/s10559-021-00368-4
13. Kondruk N., Malyar M. Dimensionality Reduction of the Criterion Space in Some Optimization Problems. International Conference "Computational Intelligence", 28–30 September, Kyiv-Uzhhorod, Ukraine, 2021. Pp. 112–121. DOI: http://ceur-ws.org/Vol-3018/Paper_11.pdf

Kondruk N. E. Models of multivariate forecasting.

This study is a development of applied data analysis. Information technologies of multifactor forecasting based on MLR and DR models are presented. An information-analytical system in the Python programming language and the scikit-learn library has been developed, which implements the descriptions of the approach. As an approbation model, the current task of forecasting Ukraine's GDP by indicators: inflation index, population, official dollar exchange rate, unemployment rate and migration growth was chosen. The training sample contained 16 observations. The regression tree model is adjusted with a coefficient of determination of 99% and an average absolute percentage error of 6%. These

quality indices of the model show its high accuracy. These quality indices of the model indicate its high accuracy.

Keywords: multivariate forecasting, regression trees, multivariate linear analysis, GDP forecasting.

References

1. Gogtay, N. J., Deshpande, S. P., & Thatte, U. M. (2017). Principles of regression analysis. *Journal of the Association of Physicians of India*, 65(48), 48–52.
2. Daoud, J. I. (2017, December). Multicollinearity and regression analysis. *In Journal of Physics: Conference Series*, 949(1), p. 012009. IOP Publishing. https://doi.org/10.1088/1742-6596/949/1/012009
3. Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8(2), 39–42. https://doi.org/10.12691/ajams-8-2-1
4. Torgo, L. (2017). Regression Trees. In: Sammut C., Webb G. I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_717
5. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification And Regression Trees (1st ed.). *Routledge*. https://doi.org/10.1201/9781315139470
6. Ministry of Finance of Ukraine. Retrieved from https://minfin.com.ua/ua/
7. Ayyadevara, V. K. (2018). Linear regression. *In Pro Machine Learning Algorithms*, 17–47. Apress, Berkeley, CA. https://doi.org/10.1016/j.eswa.2017.04.003
8. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130. https://doi.org/10.11919/j.issn.1002-0829.215044
9. State Statistics Service of Ukraine. Retrieved from http://www.ukrstat.gov.ua/
10. Kondruk, N. E. (2021). Use of similarity measures in classification methods. *Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics*, 1(38), 85–91. https://doi.org/10.24144/2616-7700.2021.38(1).143-148
11. Kondruk, N. E. (2018). Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control*, 3(46), 98–105. https://doi.org/10.15588/1607-3274-2018-3-11
12. Kondruk, N. E., & Malyar, M. M. (2021). Analysis of Cluster Structures by Different Similarity Measures. *Cybern Syst Anal*, 57, 436–441. https://doi.org/10.1007/s10559-021-00368-4
13. Kondruk, N., & Malyar, M. (2021). Dimensionality Reduction of the Criterion Space in Some Optimization Problems, 112–121. http://ceur-ws.org/Vol-3018/Paper_11.pdf

Одержано 15.04.2022