

Юрій В. Пугач*

СУЧАСНІ ЕКОНОМЕТРИЧНІ ПІДХОДИ МАШИННОГО НАВЧАННЯ У МОДЕЛЮВАННІ ФАКТОРІВ ВПЛИВУ НА ВИДАТКИ МІСЦЕВИХ БЮДЖЕТІВ

У статті запропоновано розширити підходи щодо процесів планування та бюджетування. Використано технології машинного навчання на основі алгоритму розумного градієнтного бустингу дерев рішень (XGBoost Tree). Надійні оцінки видатків забезпечили зв'язок між різними соціально-економічними характеристиками регіонів та дали змогу спрогнозувати рівень впливу факторів на витрати місцевих бюджетів.

Ключові слова: Видатки місцевих бюджетів, фінансова децентралізація, економетрика, машинне навчання.

Табл. 1. Рис. 2. Літ. 9.

DOI: 10.32752/1993-6788-2023-1-261-67-74

Yurii V. Puhach

MODERN ECONOMETRIC APPROACHES OF MACHINE LEARNING IN THE SIMULATION OF FACTORS INFLUENCING LOCAL BUDGET EXPENDITURES

The article proposes to expand approaches to planning and budgeting processes. Machine learning technologies were used based on the intelligent gradient boosting algorithm of decision trees (XGBoost Tree). Reliable estimates of expenditures provided a connection between different socio-economic characteristics of regions and made it possible to predict the level of influence of factors on the expenditures of local budgets.

Keywords: Expenditures of local budgets, financial decentralization, econometrics, machine learning.

Постановка проблеми. В умовах трансформаційних змін державного сектору, питання ефективної діагностики бюджетного потенціалу набувають все більшого значення.

Із запровадженням реформи децентралізації постає питання пошуку та наукового обґрунтування перспективної методології виявлення зв'язків та аналізу факторів впливу на статті місцевих бюджетів.

Визначення справедливих регіональних потреб є надзвичайно актуальним завданням. Сьогодні потреби у видатках формуються здебільшого за історичним принципом. Поступове зміщення пріоритетів від перерозподілу ресурсів до надання стимулів для вирішення потреб з посиленням спроможності залучати місцеві доходи має стати твердим напрямком руху центрального уряду задля самодостатності громад. Зосередження на розробці моделей оцінки витрат та їх цінності має ґрунтуватися на достовірних даних, а головне — враховувати можливі ризики кліматичних змін, умов господарювання в регіоні, наслідки від пандемій, потенційні міграційні процеси, можливі військові загрози тощо.

* Sumy National Agrarian University, Sumy

Аналіз останніх досліджень і публікацій. Проблематику регіонального розвитку у розрізі визначення справедливих регіональних потреб досліджували велика кількість науковців. Зарубіжні та вітчизняні вчені за допомогою економетричного інструментарію поглиблюють дослідження взаємозв'язків економічних процесів зі сталим розвитком країни. Серед них С. Слухай [1], І. Пишнограєв, Л. Горошкова [2], Є. Хлобистов [3], А. Козер, К. Албу [4], Дж. Ван, В. Жун, Д. Мей [5] та інші. Використання машинного навчання у виявленні зв'язків між економічними процесами останнім часом набуває затребуваності.

Проте напрацювань, адаптованих до українських реалій в умовах розширення фінансової децентралізації недостатньо, а тому дане питання потребує подальших досліджень.

Метою дослідження є поглиблення теоретико-методологічних засад та обґрунтування методологічного інструментарію можливостей машинного навчання у моделюванні факторів впливу на видатки місцевих бюджетів.

Основні результати дослідження. Діагностика та прогнозування розвитку адміністративно-територіальних одиниць, оцінка фінансової спроможності за допомогою сучасних економетричних підходів убачається важливим та актуальним напрямком досліджень. Задля досягнення поставлених цілей місцеві рівні врядування повинні мати змогу оцінювати фінансові потреби на перспективу: визначати можливі варіанти мобілізації ресурсів, узгоджувати стратегії розвитку з плануванням потреб у видатках, пропонувати раціональний розподіл державних фінансів, розробляти умови взаємодії з партнерами з недержавного сектору, відпрацьовувати напрямки інвестиційних надходжень, посилювати підзвітність та прозорість розподілу ресурсів. Належне прогнозування витрат має вирішальне значення для стійкості фіскальної програми і підготовки бюджету, особливо у багаторічному фінансовому плануванні.

У розвинених країнах кілька десятиліть тому з початком застосування новітніх підходів до моделювання економічних процесів технології машинного навчання мали значний вплив на бізнес, бюджетування доходів. Для багатьох організацій це стало важливим інструментом для самовдосконалення, розпізнавання певних завдань та прогнозування. Підвищений інтерес до такого підходу пов'язаний з можливістю долучення безлічі факторів, включаючи сценарії обробки об'ємних даних за розумних витрат та за відносно короткий час виконання.

Для визначення напрямків покращення та удосконалення рівня фінансової децентралізації за допомогою контролю видаткової частини місцевих бюджетів нами було запропоновано сучасний підхід машинного навчання – гнучкої високоефективної моделі XGBoost, в основі якої лежить алгоритм градієнтного бустингу дерев рішень.

Запровадження зазначеного підходу для визначення фінансових потреб на рівні країни викликано необхідністю забезпечення методологією з урахуванням ризиків та залученням зацікавлених сторін. Долучення міністерств, місцевих органів влади, представників наукових кіл до осучаснення підходів у розрахунках та визначенні витрат місцевих бюджетів

повинно поглибити підходи та сприяти ефективному використанню фінансових ресурсів.

На початковому етапі застосування слід структурувати систему специфічних індикаторів, які послужать основою потужної підтримки моніторингу та оцінки місцевого прогресу у досягненні результатів розвитку.

За допомогою певних класифікацій алгоритм намагається сам знайти приховані закономірності в даних і вчитися на цьому. Згідно з цими підходами, потреби у видатках місцевого самоврядування визначаються на основі статистичних зв'язків між характеристиками регіону, обраними як «рушійної сили потреб» (вікова структура, чисельність населення тощо) та минулими доходами. Такі оцінені співвідношення використовуються у формулах для розрахунку потреб у видатках місцевих рад. Надійні оцінки витрат можуть забезпечити розуміння наявного фінансування, оцінити фінансові ризики, оцінити стабільність послуг, визначити ключові змінні, що визначають зміну рівня витрат. Зрештою, такий підхід дає розуміння впливу фінансової децентралізації на державні витрати місцевих органів влади. Органи місцевого самоврядування приділяють все більше уваги тенденціям витрат і їх точні оцінки можуть допомогти керувати бюджетною рівновагою та підвищити довіру громадян.

Перш ніж залучати до моделі показники необхідно прийняти компромісне рішення між простотою та точністю. Адже формула лише з кількома факторами впливу на потреби зменшує складність розрахунків та буде легшою для розуміння, тоді як формула з додатковими показниками буде більш точною. Ця дилема вирішується за допомогою застосування потужних програмних комплексів, які забезпечують швидкість обробки інформації, гнучкість підходів, точність розрахунків з використанням великого об'єму вхідних даних. Тому для подальшої роботи з даними обрано IBM SPSS Modeler V.18.3.0 [6], як краще програмне забезпечення 2022 року в категорії продуктів Analytics і AI [7]. Методи, доступні в палітрі моделювання, дозволяють отримувати нову інформацію з даних, виявляти раніше приховані закономірності і розробляти прогностичні моделі швидко та інтуїтивно без безпосереднього програмування.

Для демонстративного моделювання було використано набір статистичної інформації, яка характеризує зміни в регіонах. Модель може доповнюватися аналітичними характеристиками областей, враховуючи якісні (двійкові, категоріальні дані) та кількісні показники.

Завантаживши дані до програми перш за все необхідно виконати їх обробку – відфільтрувати. Для коректного застосування змінних визначаємо тип даних в залежності від їх значень та ролі у моделі (вхідний чи цільовий предиктор). Програму варто примусити перевіряти значення і зазначити комірки без даних як «white space», щоб в подальшому відсутність інформації не інтерпретувалася як нульові значення.

Окрім цього штучним шляхом визначимо відсутні значення цих показників за допомогою запропонованого згенерованого програмою автоматичного алгоритму прогнозування рішень, а саме модуля інтерактивних дерев (C&RT, CHAID) для збагачення інтерпретації.

Підготовані дані за допомогою різновидів графіків, виводу аналізованої інформації було перевірено та досліджено на розподіл значень, їхню якість.

Згенерувавши нові дані ми отримали 100% наповнюваність, що сприятиме якісному аналізу в подальшому, хоча при цьому може постраждати коефіцієнт кореляції визначених предикторів за рахунок додавання розрахованих даних.

Щоб уникнути переобладнання моделі та оптимізувати для подальшого аналізу, дані поділяються на навчальний і тестовий набори. Алгоритм використовує навчальний набір даних для створення моделі та тестовий набір – для вимірювання продуктивності. Чим більше даних використовується як навчальні, тим краще модель зможе вчитися на основі даних і модель краще відображає реальність. При цьому тестовий набір буде меншим, що призводить до меншої впевненості в результатах, оскільки продуктивність вимірюється на меншій кількості записів. Поширеною і часто ефективною практикою для машинного навчання є практика 80/20. 80% даних використовується для навчання, а 20% даних використовується для тестування та оцінки побудованої моделі. Це гарантує, що модель не надто застосована до навчальних даних і може узагальнювати нові дані. Вибірковість визначення рядків для присвоєння типу набору визначаємо випадковим генеруванням. Чітка прив'язка до розподілу вибірки зумовлена ще тим, що наше завдання пов'язане з прогнозуванням значень видатків місцевих бюджетів. Визначаємо тип згенерованих даних.

У поєднанні великої кількості даних та потужності обчислювальної техніки скористаємося сучасним та одним із найефективніших методів машинного навчання XGBoost (eXtreme Gradient Boosting) [8], [9]. Найважливішими особливостями алгоритму є те, що він може досягати високої потужності прогнозування, запобігати перенавчанню, керувати пустими даними та вираховувати їх, працює в 10 разів швидше ніж інші популярні алгоритми.

Така модель приймає інтегровану ідею і може використовуватися для вирішення проблем як класифікації, так і регресії. Алгоритм в основному застосовує інтегральні обчислення, вирішує функцію мінімальних втрат за допомогою розкладання Тейлора другого порядку, визначає розділений вузол і буде остаточною моделлю.

Даний вузол на платформі SPSS Modeler в доступній формі розкриває основні функції та часто використовувані параметри цієї моделі. В основі XGBoost Tree лежить алгоритм розумного градієнтного бустингу дерев рішень. Особливістю XGBoost Tree є побудова моделі підвищення ефективності на основі ітеративного вивчення слабких класифікаторів з подальшим додаванням їх до остаточного сильного класифікатора. Дозволяє навчитися нової моделі на підставі даних попередньої, а також заздалегідь розрахованих відхилень прогнозу від факту [4].

Моделю XGBoost в основному включає три типи параметрів: загальні параметри (для встановлення загальних функцій моделі), параметри просування (для швидкості навчання) та параметри навчання (для керування виконанням оптимізації завдання).

На основі цих трьох типів параметрів ми оптимізували важливі параметри та створили оптимізовану модель, щоб покращити прогностний вплив моделі на видатки місцевих бюджетів.

Для успішної реалізації алгоритму XGBoost Tree перейдемо безпосередньо до формулювання параметрів для побудови моделі.

Аналізуємо поля, які ми будемо використовувати та призначаємо їм заздалегідь визначені ролі. Нашою метою є дослідження показника видатків місцевих бюджетів (EXPLOCBUDG) як цільової функції. Всі інші предиктори будуть виступати у якості вхідних змінних для прогнозування.

Визначаємо основні гіперпараметри нашої моделі вручну, щоб прогноз адаптувався якомога ближче до свого справжнього значення. В подальшому порівнюємо показники після використання гіперпараметрів за замовчуванням та оптимізованих програмним комплексом.

На основі вищенаведених дій з основними налаштуваннями ми змогли визначити, що результати прогнозування класифікації за допомогою тестової моделі з користувацькими налаштуваннями були кращими, ніж результати моделей за замовчуванням та оптимізованої автоматично програмним комплексом (табл. 1). Тому для подальшого аналізу та використання для прогнозування будемо керуватися саме цим варіантом відбору параметрів.

Таблиця 1. Порівняльна характеристика вихідних параметрів моделей XGBoost за ступенем автоматизації, розроблено автором

Показник	Гіперпараметри встановлені за замовчуванням		Гіперпараметри оптимізовані програмою		Гіперпараметри оптимізовані вручну	
	Навчання	Тест	Навчання	Тест	Навчання	Тест
Mean Absolute Error	725,886	1730,07	2060,43	1761,48	752,93	1477,56
Standard Deviation	933,57	2725,07	2046,611	2005,07	999,74	2258,96
Linear Correlation	0,998	0,975	0,995	0,97	0,998	0,98
Occurrences	120	30	120	30	120	30

Ми використали алгоритм XGBoost для виконання обчислень в програмному забезпеченні IBM SPSS Modeler на базі мови програмування Python, щоб отримати оцінки важливості ознак для компонентних факторів [5]. З рисунку 1 ми бачимо розташування основних компонентів, які були задіяні в моделі. Поєднання надійного теоретичного обґрунтування та швидкого практичного алгоритму робить використані предиктори потужним інструментом для впевненої інтерпретації деревоподібної моделі підвищення градієнта XGBoost.

З гістограми можемо зробити певні висновки - визначити важливість функції для кожної змінної у нашому наборі даних. Задіяно було в побудові моделі 28 із 36 предикторів. Із них F2 – 11, F13 – 9, F1 – 7, F4 – 6 з відповідними значеннями мають найбільший внесок у точне прогнозування.

Тобто від розміру міжбюджетних трансфертів з державного бюджету (F2), зібраних податків на доходи фізичних осіб (F13), єдиного податку (F1),

наявного доходу населення (F4) найбільше залежать на даному етапі розвитку місцевого самоврядування прогностні показники видатків з місцевих бюджетів, як важливого чинника впливу на рівень фінансової децентралізації.

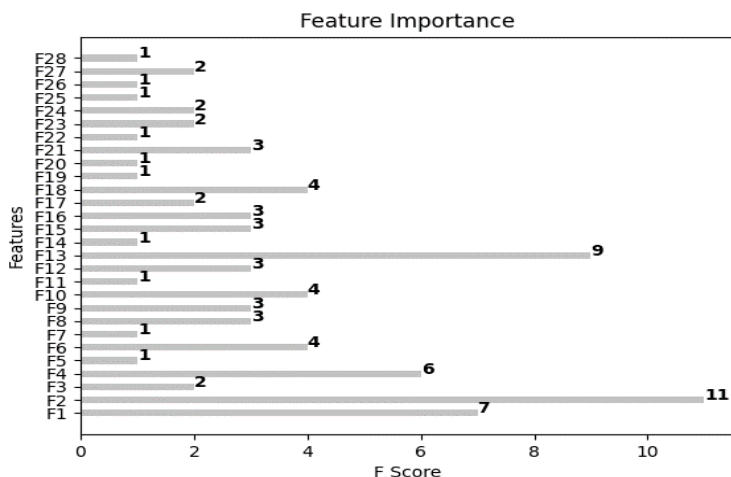


Рис. 1. Оцінка важливості предикторів моделі XGBoost, побудовано автором за допомогою програмного комплексу SPSS Modeler

У той же час оцінки інших показників, таких як розмір імпорту (F6), кількість зайнятого населення в промисловості (F10), місцеві податки (F18), поточні витрати на охорону НПС (F8), кількість міського населення (F9), розмір капітальних інвестицій (F12), кількість зайнятого населення в будівництві (F15), розмір експорту (F16) також не були низькими. Це вказує на те, що ці основні компоненти також певною мірою вносять свій внесок у прогнозування видаткової частини місцевих бюджетів. Решта показників використовувались моделлю меншу кількість разів, а отже мають несуттєвий вплив на цільовий показник. Натомість програмою не були задіяні жодного разу у побудові дерева показники капітальних інвестицій на охорону навколишнього природного середовища, міграційного приросту населення, зайнятого населення у транспортній, поштової, професійній, науковій, технічній, іншого виду економічної діяльності, населення задіяне в охороні здоров'я та наданні соціальної допомоги, працівники задіяні у виконанні наукових досліджень і розробок зі складу допоміжного персоналу, кількості проєктів публічно-приватного партнерства.

Модель має не тільки якісний прогностний вплив на навчальний набір, а й на тестовий набір, що вказує на високу здатність до узагальнення. Крива прогнозу навчальної моделі майже співпадає з кривою ідеальної моделі прогнозу (Рис. 2). Крива \$BEST-EXPLOCBUDG знаходиться на певному відрізку над кривою \$XGT- EXPLOCBUDG, що вказує на покращення в якості моделі на цьому відрізку навчання.

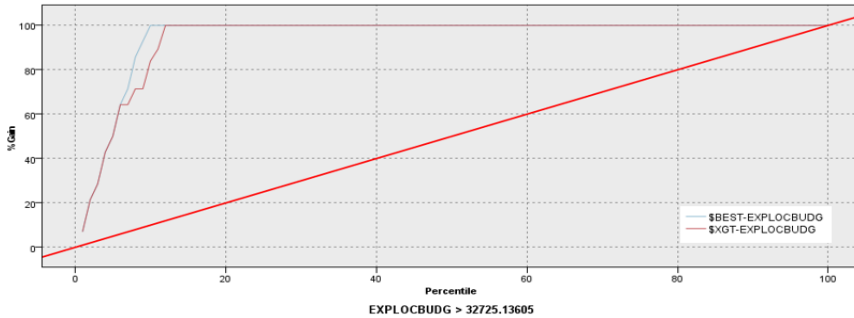


Рис. 2. Зміна точності кривих покращення моделі при збільшенні кількості дерев рішень, побудовано автором за допомогою програмного комплексу SPSS Modeler

Графік кривих навчання надає діагностичний інструмент, який можна інтерпретувати та пропонувати конкретні зміни гіперпараметрів моделі у разі низької ефективності прогнозування. Побудована модель актуальною, стійкою, є доречною з точки зору прогнозування. Аналіз значень дав змогу оцінити продуктивність моделі на кожній ітерації алгоритму під час навчання та побудувати результати у вигляді кривих навчання.

Таким чином, ми використали метод екстремального підвищення градієнта XGBoost, що дало змогу встановити зв'язок між різними соціально-економічними характеристиками регіонів та спрогнозувати рівень впливу факторів на витрати місцевих бюджетів. Слід зазначити, що запропонований метод дав змогу досягти високих показників ефективності, що є доказом дієвості моделі.

Висновки. Розроблено та використано сучасні економетричні підходи машинного навчання у моделюванні факторів впливу на видатки місцевих бюджетів. Застосовано підхід до моделювання економічних процесів за допомогою технології машинного навчання на основі алгоритму розумного градієнтного бустингу дерев рішень (XGBoost Tree). Надійні оцінки видатків забезпечили розуміння наявного фінансування, дали можливість оцінити фінансові ризики, стабільність послуг, визначити ключові змінні, що визначають зміну рівня видатків. Такий підхід дав змогу встановити зв'язок між різними соціально-економічними характеристиками регіонів та спрогнозувати рівень впливу факторів на витрати місцевих бюджетів. Обґрунтовано базову модель оцінювання зв'язків видаткової частини місцевих бюджетів на основі технології машинного навчання, доведено практичну значущість.

Це дослідження визначає один із способів удосконалення поточних підходів фінансування. Важливість знань про сучасні економетричні підходи має стати запорукою успішної реалізації реформи децентралізації та оптимального розподілу фінансових ресурсів.

1. Слухай С. Вплив децентралізації на економіку регіонів України. Електронне фахове наукове видання «Ефективна економіка». 2022. №1. С.1-5. DOI: 10.32702/2307-2105-2022.1.15

2. Пишнограєв І. Моделювання впливу соціально-економічних кризових явищ на міграційні процеси. Проблеми системного підходу в економіці. 2021. №6(86). С. 94-102. DOI: <https://doi.org/10.32782/2520-2200/2021-6-13>

3. Горошкова Л., Лісовський С., Хлобистов Є. Потенціал ефективності реформування адміністративно-територіального устрою (на прикладі Запорізької області). Проблеми безперервної географічної освіти і картографії. 2019. №30. С. 32-42. DOI: 10.26565/2075-1893-2019-30-04

4. Co er A., Maer-matei M. та Albu C. Predictive models for loan default risk assessment. Economic Computation & Economic Cybernetics Studies & Research. 2019. v. 53, № 2, pp. 149-165. DOI:10.24818/18423264/53.2.19.09

5. Wang J., Rong W., Zhang Z., and Mei D. Credit Debt Default Risk Assessment Based on the XGBoost Algorithm: An Empirical Study from China. Wireless Communications and Mobile Computing. 2022. № 8, pp. 1-14. DOI: <https://doi.org/10.1155/2022/8005493>

6. IBM Corporation. IBM SPSS Modeler 18.3 Applications Guide. веб-сайт URL:https://www.ibm.com/docs/en/SS3RA7_18.3.0/pdf/ModelerApplications.pdf.

7. G2. Best Software by Function / Analytics and AI Products. G2.com, Inc., 2022. веб-сайт URL: <https://www.g2.com/best-software-companies/top-analytics-and-ai#list>.

8. Chen T. та Guestrin C. XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 785-794. DOI: <https://doi.org/10.1145/2939672.2939785>

9. Li C., Zheng X., Yang Z. та Kuang L. Predicting Short-Term Electricity Demand by Combining the Advantages of ARMA and XGBoost in Fog Computing Environment. Hindawi, Wireless Communications and Mobile Computing. 2018. № 8, pp. 1-18. DOI:10.1155/2018/5018053

1. Slukhai S. Vplyv detsentralizatsii na ekonomiku rehioniv Ukrainy. Elektronne fakhove naukove vydannia «Efektyvna ekonomika». 2022. №1. S.1-5. DOI: 10.32702/2307-2105-2022.1.15

2. Pyshnohraiev I. Modeliuvannia vplyvu sotsialno-ekonomichnykh kryzovykh yavysch na mihrat-siini protsesy. Problemy systemnoho pidkhotu v ekonomitsi. 2021. №6(86). S. 94-102. DOI: <https://doi.org/10.32782/2520-2200/2021-6-13>

3. Horoshkova L., Lisovskyi S., Khlobystov Ye. Potentsial efektyvnosti reformuvannia administratyvno-terytorialnoho ustroiu (na prykladi Zaporizkoi oblasti). Problemy bezperervnoi heohrafichnoi osvity i kartohrafi. 2019. №30. S. 32-42. DOI: 10.26565/2075-1893-2019-30-04

4. Co er A., Maer-matei M. та Albu C. Predictive models for loan default risk assessment. Economic Computation & Economic Cybernetics Studies & Research. 2019. v. 53, № 2, pp. 149-165. DOI:10.24818/18423264/53.2.19.09

5. Wang J., Rong W., Zhang Z., and Mei D. Credit Debt Default Risk Assessment Based on the XGBoost Algorithm: An Empirical Study from China. Wireless Communications and Mobile Computing. 2022. № 8, pp. 1-14. DOI: <https://doi.org/10.1155/2022/8005493>

6. IBM Corporation. IBM SPSS Modeler 18.3 Applications Guide. веб-сайт URL:https://www.ibm.com/docs/en/SS3RA7_18.3.0/pdf/ModelerApplications.pdf.

7. G2. Best Software by Function / Analytics and AI Products. G2.com, Inc., 2022. веб-сайт URL: <https://www.g2.com/best-software-companies/top-analytics-and-ai#list>.

8. Chen T. та Guestrin C. XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 785-794. DOI: <https://doi.org/10.1145/2939672.2939785>

9. Li C., Zheng X., Yang Z. та Kuang L. Predicting Short-Term Electricity Demand by Combining the Advantages of ARMA and XGBoost in Fog Computing Environment. Hindawi, Wireless Communications and Mobile Computing. 2018. № 8, pp. 1-18. DOI:10.1155/2018/5018053