УДК 519.86:004.8:332.6

Ю. В. КЛЕБАН, Л. О. ЛЕПСЬКА
Національний університет «Острозька академія»

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА АНАЛІЗ ВАЖЛИВОСТІ ОЗНАК У ЗАДАЧІ ПРОГНОЗУВАННЯ ЦІН НА ВТОРИННОМУ РИНКУ ЖИТЛА З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ

У статті досліджено задачу прогнозування цін на вторинному ринку житлової нерухомості м. Києва шляхом побудови та порівняння альтернативних регресійних моделей. Актуальність дослідження визначається багатофакторною природою формування вартості житла, високою волатильністю ринку та необхідністю підвищення точності аналітичних інструментів оцінювання. Класичні параметричні підходи не завжди дозволяють адекватно врахувати нелінійні залежності та взаємодії між характеристиками об'єктів, що обґрунтовує використання сучасних алгоритмічних методів.

У дослідженні реалізовано комплексну процедуру підготовки даних, що включала очищення вибірки, імпутацію пропущених значень методом k-найближчих сусідів, вінзоризацію екстремальних спостережень, логарифмічне перетворення цільових змінних та нормалізацію числових ознак. Особливу увагу приділено інженерії ознак на основі текстових описів об'єктів. Повторювані якісні характеристики було ідентифіковано та закодовано у вигляді бінарних змінних, що дозволило включити інфраструктурні параметри та показники комфорту до аналітичної структури. Додатково здійснено геопросторове збагачення даних шляхом розрахунку відстані кожного об'єкта до центру міста як ключового індикатора локації.

Побудовано та порівняно моделі множинної лінійної регресії, ансамблеві алгоритми (XGBoost, LightGBM, Random Forest) та штучну нейронну мережу. Оцінювання якості моделей здійснювалося на тестовому наборі даних (20 % спостережень) з використанням метрик R^2, RMSE, MAE та MAPE. Найвищу прогнозну точність продемонстрували алгоритми бустингу, зокрема LightGBM.

Аналіз важливості ознак показав, що загальна площа є домінуючим чинником формування загальної вартості житла разом із локаційними характеристиками та віком будинку. Натомість ціна за квадратний метр сильніше залежить від параметрів розташування та адміністративної приналежності. Отримані результати підтверджують ефективність ансамблевих алгоритмів у поєднанні з розширеною інженерією ознак для кількісного аналізу механізмів ціноутворення на місцевому ринку житлової нерухомості.

Ключові слова: *регресійне моделювання, ансамблеві алгоритми, аналіз важливості ознак, інженерія ознак, вторинний ринок житлової нерухомості, прогнозування цін.*

Yu. V. KLEBAN, L. O. LEPSKA
The National University of Ostroh Academy

MATHEMATICAL MODELING AND FEATURE IMPORTANCE ANALYSIS IN THE PROBLEM OF HOUSING PRICE FORECASTING IN THE SECONDARY MARKET USING MACHINE LEARNING

The paper investigates the problem of price prediction in the secondary residential housing market of Kyiv through the construction and comparison of alternative regression models. The relevance of the study is determined by the multifactor nature of housing price formation, high market volatility, and the need to improve the accuracy of analytical valuation tools. Classical parametric approaches do not always adequately capture nonlinear relationships and interactions between property characteristics, which justifies the use of modern algorithmic methods.

A comprehensive data preparation procedure was implemented, including data cleaning, missing value imputation using the k-nearest neighbors algorithm, Winsorization of extreme observations, logarithmic transformation of target variables, and normalization of numerical features. Particular attention was given to feature engineering based on textual property descriptions. Recurrent qualitative characteristics were identified and encoded as binary variables, enabling the incorporation of infrastructural and comfort-related attributes into the analytical framework. Additionally, geospatial enrichment was performed by calculating the distance of each property to the city center as a key location indicator.

Multiple linear regression, ensemble algorithms (XGBoost, LightGBM, Random Forest), and an artificial neural network were developed and compared. Model performance was evaluated on a test dataset (20 % of observations) using R^2, RMSE, MAE, and MAPE metrics. Boosting algorithms, particularly LightGBM, demonstrated the highest predictive accuracy.

Feature importance analysis revealed that total area is the dominant determinant of overall housing price, together with location-related characteristics and building age. In contrast, price per square meter is more strongly driven by loca-

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

tional parameters and administrative affiliation. The results confirm the effectiveness of ensemble algorithms combined with extended feature engineering for quantitative analysis of price formation mechanisms in the local residential real estate market.

Keywords: *regression modeling, ensemble algorithms, feature importance analysis, feature engineering, secondary housing market, price prediction.*

Постановка проблеми

Ринок житлової нерухомості є важливим сегментом економіки України, а прозорість і обґрунтованість формування ринкової вартості житла мають суттєве значення для фінансової стабільності. Згідно зі «Звітом про фінансову стабільність» Національного банку України (грудень 2025 р.), попри поступове відновлення ринку, інформаційна асиметрія та коливання попиту ускладнюють об'єктивну оцінку активів [1]. Це зумовлює потребу у більш точних аналітичних інструментах прогнозування.

На нормативному рівні автоматизація оцінки нерухомості закріплена Постановою Кабінету Міністрів України від 23 червня 2021 р. № 681 щодо функціонування Єдиної державної електронної системи у сфері будівництва [2], а також положеннями Закону України «Про оцінку майна, майнових прав та професійну оціночну діяльність в Україні» [3], що передбачають розвиток автоматизованих модулів оцінки.

Водночас традиційні економетричні моделі не завжди адекватно відображають складні нелінійні залежності між характеристиками об'єктів та їхньою ринковою вартістю. Ціна житла формується під впливом числових і текстових ознак, що обґрунтовує застосування методів машинного навчання та включення текстових характеристик оголошень до вектору ознак моделі з метою підвищення точності прогнозування.

Аналіз останніх досліджень і публікацій

Дослідження ринку нерухомості традиційно ґрунтуються на моделях, у яких вартість об'єкта розглядається як функція його фізичних і локаційних характеристик. У межах таких підходів оцінювання здійснюється за допомогою регресійного аналізу, що передбачає переважно лінійну залежність між параметрами житла та його ціною. Подібну методологію представлено, зокрема, у О. Пашкевич, С. Ващищак, А. Бойчук та ін. [4]. Водночас, як зазначають Ходжа В. та Шала А. [5], за умов високої волатильності та структурної мінливості ринку лінійні економетричні моделі можуть демонструвати обмежену прогностичну стійкість через нелінійний характер взаємозв'язків між факторами.

У сучасних дослідженнях дедалі ширше застосовуються методи машинного навчання, які дозволяють моделювати складні багатофакторні залежності без жорсткого припущення про форму функціонального зв'язку. В оглядовій роботі Морено-Форонда І. та ін. [6] показано, що ансамблеві та нелінійні алгоритми забезпечують вищу точність прогнозування порівняно з класичними регресійними моделями. Подібні результати наведено в дослідженні Алкан Т. та ін. [7], де підкреслено здатність методів Soft Computing адаптуватися до багатовимірної структури даних, характерної для міських ринків нерухомості.

Порівняльна оцінка алгоритмів машинного навчання також отримала подальший розвиток у наукових працях. У роботі Яздані М. [8] продемонстровано переваги деревоподібних моделей і методів глибокого навчання над лінійною регресією у задачах прогнозування вартості житла. Дослідження Джа С. Б. та ін. [9] підтверджує здатність таких алгоритмів виявляти приховані закономірності ціноутворення на основі емпіричних даних. В українських наукових роботах методичні аспекти підготовки даних та алгоритмізації процесу моделювання розглянуто у працях Вереса О. та Шимоняка А. [10].

Окрему групу досліджень становлять роботи, присвячені використанню неструктурованих даних. Чжан Х., Лі Ю. та Бранко П. показали, що включення текстових описів оголошень до моделей прогнозування забезпечує статистично значуще підвищення точності оцінювання

[11]. Автори порівняли моделі, побудовані лише на числових ознаках, лише на текстових характеристиках та на їх поєднанні, і встановили перевагу комбінованого підходу. Для представлення текстових даних у їх дослідженні застосовано методи векторизації (TF-IDF, Word2Vec, BERT), що дозволяють враховувати семантичні особливості описів об'єктів.

Попри наявність значної кількості досліджень, питання поєднання бустингових алгоритмів (зокрема XGBoost і LightGBM) із включенням текстових характеристик оголошень у вигляді додаткових структурованих ознак для моделювання вторинного ринку нерухомості Києва залишається недостатньо висвітленим у науковій літературі. Більшість робіт зосереджуються або на структурованих кількісних показниках, або на окремому використанні методів глибокої текстової обробки. Крім того, обмежено досліджено роздільне моделювання загальної вартості об'єкта та ціни за квадратний метр в умовах українського ринку 2025–2026 років, що обґрунтовує актуальність проведеного дослідження.

Мета дослідження

Метою дослідження є визначення та кількісна оцінка факторів, що впливають на ціну житлової нерухомості, із застосуванням методів машинного навчання та розширенням набору предикторів за рахунок текстових характеристик оголошень. Дослідження передбачає порівняння класичних статистичних моделей і сучасних алгоритмів ML з метою встановлення найбільш значущих ознак і підвищення точності прогнозування. Особливістю роботи є роздільне моделювання загальної вартості об'єкта та ціни за квадратний метр, що дозволяє коректніше враховувати вплив площі та структурних характеристик житла.

Виклад основного матеріалу дослідження

Обґрунтування вибору інструментів та метрик. Прогнозування вартості житлової нерухомості належить до задач регресійного моделювання, у межах яких ціна об'єкта розглядається як функція його характеристик. На формування вартості впливають фізичні, технічні та локаційні параметри житла, що зумовлює багатовимірний характер задачі. Класичні економіко-математичні методи, зокрема лінійна регресія, дозволяють оцінити внесок окремих факторів, однак їх застосування пов'язане з припущеннями щодо форми функціональної залежності та статистичних властивостей залишків, що обмежує точність моделювання у разі наявності нелінійних зв'язків.

Для врахування складних взаємодій між ознаками застосовано алгоритми машинного навчання, зокрема XGBoost, LightGBM, Random Forest та багатошарові нейронні мережі. Бустингові алгоритми (XGBoost, LightGBM) реалізують послідовне побудування ансамблю дерев рішень із мінімізацією функції втрат, Random Forest формує агрегований прогноз на основі незалежних дерев, а нейронні мережі апроксимують нелінійні залежності у багатовимірному просторі ознак.

Якість моделей оцінювалась за допомогою стандартних метрик регресії: середньоквадратичної помилки (RMSE), середньої абсолютної помилки (MAE), середньої абсолютної відносної помилки (MAPE) та коефіцієнта детермінації (R^2), що забезпечує можливість порівняння їхньої прогностичної здатності.

Дослідження спрямоване на моделювання вартості житлової нерухомості Києва на основі структурованих кількісних характеристик об'єктів, а також додаткових текстових ознак, виділених із описів оголошень та закодованих у вигляді бінарних змінних. Побудовано моделі для прогнозування як загальної вартості об'єкта, так і ціни за квадратний метр. Такий підхід дозволяє окремо оцінити вплив площі та якісних характеристик житла на формування ціни.

Обчислення та побудову моделей здійснено з використанням мови програмування R [12]. Попередню обробку та аналіз даних виконано за допомогою пакетів tidyverse та data.table. Імпутацію пропущених значень реалізовано методом найближчих сусідів із використанням

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

пакета VIM. Для побудови та оцінювання моделей машинного навчання застосовано пакети caret, randomForest, xgboost та lightgbm. Аналіз важливості ознак здійснювався на основі вбудованих механізмів деревоподібних моделей, що дозволяють оцінити внесок змінних у зменшення функції втрат.

Текстові описи оголошень були проаналізовані з метою виділення додаткових характеристик, які надалі закодовано у вигляді бінарних (dummy) змінних та включено до вектору ознак моделі. Візуалізацію результатів і діагностичних графіків виконано за допомогою пакета ggplot2.

Для моделювання використано набір даних «Kyiv secondary (resale) residential real estate» (Kyrychenko, 2026) [13], що містить інформацію про квартири, виставлені на продаж у Києві. Датасет включає числові та категоріальні ознаки, зокрема площу, кількість кімнат, поверх, загальну кількість поверхів, район, відстань до метро, рік побудови, стан квартири та матеріал стін.

Перед навчанням моделей виконано обробку пропущених значень, аналіз викидів і кодування категоріальних змінних. Для оцінювання узагальнювальної здатності моделей застосовано розділення на тренувальну і тестову вибірки та процедуру крос-валідації.

Ефективність прогнозування цін на нерухомість значною мірою залежить від якості підготовки вхідних даних. У межах дослідження реалізовано процедуру попередньої обробки, що включала очищення вибірки, формування додаткових ознак на основі текстових описів оголошень, а також перетворення та стандартизацію числових змінних.

На початковому етапі враховано наявність змінної «currency», яка відображала валюту подання вартості об’єкта (USD або UAH). Оскільки частина оголошень містила ціну в доларах США, усі значення було приведено до єдиної грошової одиниці – гривні – шляхом конвертації за фіксованим валютним курсом. Для перерахунку використовувався курс продажу 1 USD = 43,44 грн. З метою покращення інтерпретованості результатів усі цінові показники додатково поділено на 1000, тому в подальшому аналізі вони подаються у тис. грн.

Емпірична база дослідження та підготовка даних. Емпіричне дослідження ґрунтується на вибірці вторинного ринку житлової нерухомості м. Києва, яка на початковому етапі містила 14 066 спостережень і 20 базових змінних. Дані включали цінові параметри, фізичні характеристики об’єктів (площа, поверховість, рік побудови тощо), локаційні показники, а також текстові описи оголошень [12].

Підготовка даних до моделювання здійснювалася у кілька послідовних етапів і включала статистичне очищення вибірки, формування додаткових ознак на основі текстових описів, а також перетворення та кодування змінних з метою формування аналітично придатного набору даних.

Обробка пропущених значень. Аналіз структури даних виявив понад 13 тис. пропущених значень, що зумовило необхідність застосування різних методів їх обробки залежно від типу змінної. Для оцінювання масштабу проблеми було побудовано діаграму, яка відображає лише ті змінні, що містять пропуски, із зазначенням їх відсоткової частки (рис. 1). Такий підхід дозволяє зосередити увагу на змінних із неповнотою даних без перевантаження візуалізації.

Найбільшу частку пропусків зафіксовано у змінній *complex* – 43,5 %. Відсутність значення у цьому випадку інтерпретовано як відсутність належності об’єкта до житлового комплексу, тому такі спостереження було закодовано як відсутність відповідної ознаки.

Для кількісних технічних характеристик об’єктів (*area_living, area_kitchen, house_age*), де пропуски не мали змістовного обґрунтування, застосовано алгоритм *k*-найближчих сусідів (*kNN*) з параметром $k = 5$. Імпутацію виконано на основі найбільш подібних спостережень у багатовимірному просторі ознак, що дозволило зберегти структуру розподілу змінних без істотного зміщення їх статистичних характеристик.

Після завершення імпутації та перевірки даних остаточний обсяг аналітичної вибірки становив 13 997 об’єктів, що забезпечило можливість подальшого моделювання.

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

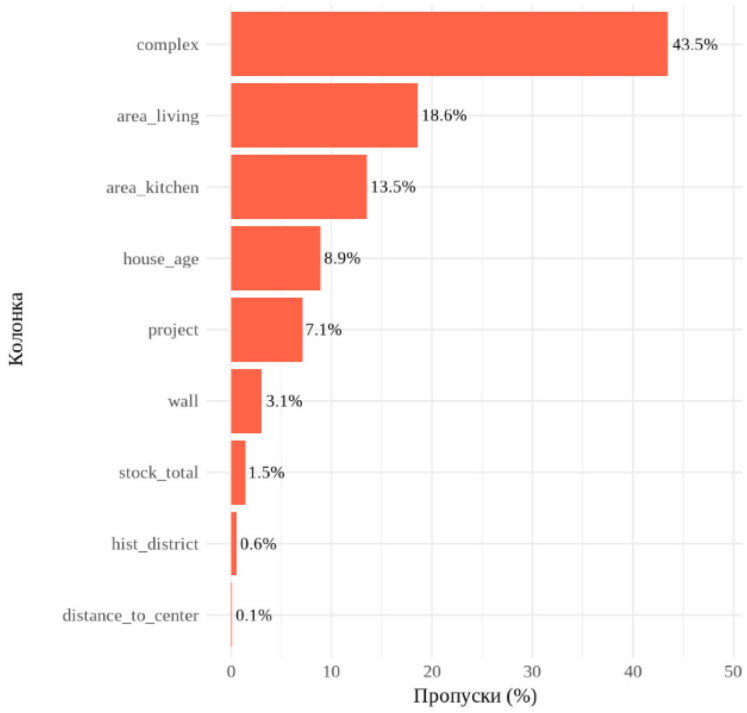

Рис. 1. Кількість пропущених значень у датасеті до обробки

Джерело: створено авторами.

Інженерія ознак на основі текстових описів. З метою розширення набору предикторів виконано формування додаткових змінних на основі текстових описів оголошень.

Для ідентифікації повторюваних характеристик у текстових описах оголошень використано інструмент ChatGPT як засіб попередньої систематизації текстових даних. До моделі було передано файл із описами квартир із досліджуваної вибірки з метою виокремлення найчастіше вживаних слів і словоформ, що відповідають окремим властивостям об'єктів. Результат було отримано у структурованому форматі JSON, який містив назву ознаки (key), перелік варіантів написання (values) та частоту їх виявлення у вибірці (count), наприклад:

```
{
«key»: «balcony»,
«values»: [«Балкон», «БАЛКОН», «балкон», «Балкони», «балкони»,
«балкончик», «лоджія», «Лоджія», «лоджию», «лоджией», «лоджии», «лоджі»],
«count»: 6099
}
```

На основі сформованого переліку відповідні характеристики було закодовано у вигляді бінарних змінних та включено до вектору ознак моделі.

Аналіз текстів дозволив ідентифікувати 13 характеристик, що відображають умови проживання та інфраструктурні особливості об'єктів – зокрема наявність охорони, ліфта, паркінгу, ремонту, побутової техніки, транспортної доступності тощо.

Виявлені текстові характеристики були закодовані у вигляді бінарних (dummy) змінних та включені до вектору ознак моделі. Перелік відповідних категорій та частота їх виявлення у вибірці наведені в табл. 1.

Аналіз частотності показав, що найбільш уживаною категорією є транспортна доступність («transport_nearby» – 9 394 згадування). Значну частоту також мають ознаки наявності балкона (6 099), школи поблизу (5 946) та елементів безпеки (4 296).

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

Таблиця 1

Перелік текстових ознак та частота їх виявлення у вибірці

№ з/п	Ознака у моделі	Варіанти ознак, що зустрічаються у текстах оголошень	Заг. к-ть повторів усіх варіантів
1	balcony	Балкон, БАЛКОН, балкон, Балкони, балкони, балкончик, лоджія, Лоджія, лоджию, лоджией, лоджии, лоджі	6099
2	elevator	ліфт, Ліфт, лифт, Лифт, ліфти, лифты	2987
3	parking	паркінг, Паркінг, парковка, Парковка, гараж, Гараж	3144
4	security	охорона, Охорона, охрана, Охрана, консьєрж, консьерж	4296
5	school_nearby	школа, Школа, школи, школы	5946
6	park_nearby	парк, Парк, сквер, Сквер	3758
7	renovation_type	євроремонт, Євроремонт, евроремонт, Евроремонт, після ремонту	729
8	new_building	новобудова, Новобудова, новострой, Новострой	352
9	heating_type	центральне опалення, централізоване опалення	110
10	furnished	з меблями, меблирована, меблированная	568
11	appliances	побутова техніка, бытовая техника	1809
12	transport_nearby	метро, Метро, зупинка, остановка	9394
13	playground	дитячий майданчик, Дитячий майданчик	555

Джерело: створено авторами з використанням ChatGPT на основі файлу з підготовленими описами житла, що використовується у дослідженні.

Після формування бінарних змінних первинну змінну «description» було вилучено з аналітичного масиву, оскільки її інформаційний зміст було відображено у створених текстових ознаках. Це дозволило уникнути надлишковості змінних під час подальшого моделювання.

Додатково виконано геокодування адрес об’єктів із визначенням їхніх географічних координат. На основі отриманих координат обчислено відстань кожного об’єкта до центральної частини міста. Сформовано змінну «distance_to_center», яка використовується як кількісний показник локаційної характеристики.

Змінну «year» трансформовано у показник віку будинку – «house_year», який розраховано як різницю між поточним роком та роком введення об’єкта в експлуатацію. Така трансформація забезпечує однозначну інтерпретацію показника в межах моделі.

У результаті інженерії ознак кількість змінних зросла з 20 до 30.

Обробка аномалій та трансформація розподілів. Аналіз розподілу цільових змінних – загальної вартості та ціни за квадратний метр – виявив правосторонню асиметрію, що може бути пов’язано з наявністю високовартісних об’єктів. Для зменшення асиметрії та стабілізації дисперсії застосовано логарифмічне перетворення.

З метою обмеження впливу екстремальних значень використано метод вінзоризації (Winsorization) на рівнях 5-го та 95-го перцентилів. Такий підхід передбачає заміну крайніх значень відповідними граничними рівнями без вилучення спостережень із вибірки. Додатково виконано перевірку на наявність викидів із використанням міжквартильного розмаху (IQR). Результати обробки окремих показників наведено на рис. 2 у формі коробкових діаграм до і після.

Нормалізація та підготовка до моделювання. Оскільки числові ознаки мали різні масштаби вимірювання (ціна – у тисячах або мільйонах гривень, площа – у десятках квадратних метрів, кількість кімнат – в одиницях), застосовано min–max нормалізацію з приведенням значень до інтервалу [0; 1]. Така трансформація зменшує вплив різниці масштабів ознак і забезпечує коректну роботу алгоритмів, чутливих до масштабування, зокрема нейронних мереж.

Кореляційний аналіз. Для попередньої оцінки взаємозв’язків між змінними побудовано кореляційну матрицю з візуалізацією у вигляді теплової карти (рис. 3).

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

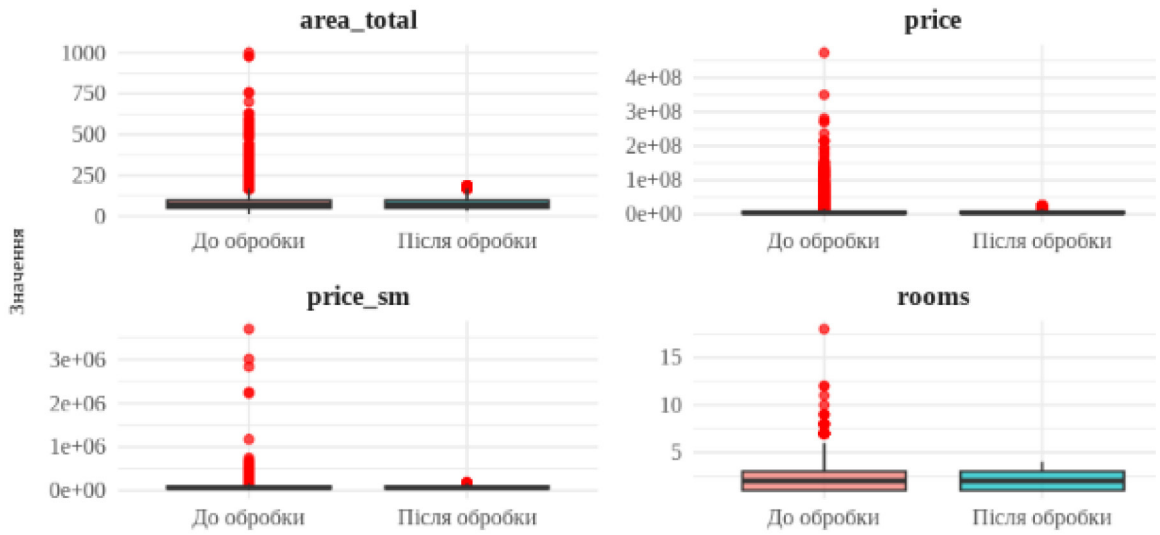

Рис. 2. Порівняння викидів до та після очищення

Джерело: створено авторами.

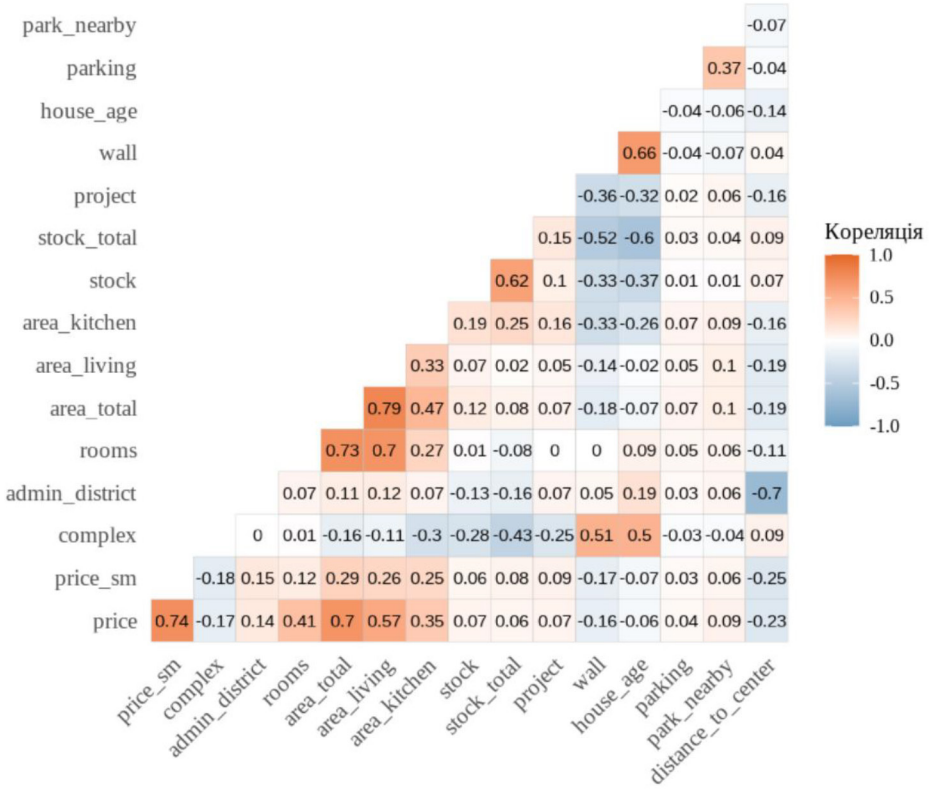

Рис. 3. Кореляційна матриця

Джерело: створено авторами.

Варто відмітити, що на рис. 3. відображено лише ті змінні, які мають коефіцієнт парної кореляції щонайменше з однією іншою змінною на рівні $|r| \geq 0{,}3$. Ознаки зі слабшими зв'язками ($|r| < 0{,}3$) не включені до візуалізації з метою підвищення інформативності графіка та концентрації уваги на найбільш суттєвих взаємозв'язках.

Проведений кореляційний аналіз показав, що найбільші значення коефіцієнта кореляції із загальною ціною житла (*price*) мають змінні, що характеризують площу об'єкта. Найвищий позитивний зв'язок спостерігається між ціною та загальною площею квартири (area_total,

$r \approx 0{,}79$). Високий рівень кореляції також зафіксовано між ціною та кількістю кімнат (rooms, $r \approx 0{,}73$), а також житловою площею (area_living, $r \approx 0{,}70$). Площа кухні (area_kitchen) демонструє помірний позитивний зв'язок із ціною ($r \approx 0{,}57$).

Ціна за квадратний метр (price_sm) має помірний позитивний зв'язок із загальною ціною ($r \approx 0{,}74$), що узгоджується зі способом її розрахунку. Водночас між price_sm та загальною площею спостерігається слабкий від'ємний зв'язок ($r \approx -0{,}18$).

Серед локаційних характеристик зафіксовано помірний негативний зв'язок між ціною та відстанню до центру міста (distance_to_center, $r \approx -0{,}23$). Подібна залежність простежується і для ціни за квадратний метр ($r \approx -0{,}25$).

Окремі якісні характеристики також демонструють статистично помітні зв'язки. Зокрема, змінна «complex» має помірний позитивний зв'язок із загальною ціною ($r \approx 0{,}41$). Поверх розташування («stock») та загальна поверховість будинку («stock_total») характеризуються помірною позитивною кореляцією між собою ($r \approx 0{,}62$), що відображає їх структурну пов'язаність.

З погляду діагностики мультиколінеарності найвищі значення коефіцієнта кореляції між незалежними змінними зафіксовано для пар area_total – rooms ($r \approx 0{,}73$), area_total – area_living ($r \approx 0{,}47$), а також stock – stock_total ($r \approx 0{,}62$). Наявність таких взаємозв'язків може вказувати на потенційну мультиколінеарність у разі одночасного включення відповідних змінних до регресійних моделей, що потребує додаткової перевірки за допомогою показників VIF або інших діагностичних критеріїв.

Таким чином, кореляційний аналіз дозволяє визначити змінні з найбільшими значеннями коефіцієнта кореляції із цільовими показниками, а також виявити структурні залежності між пояснювальними змінними, які необхідно враховувати під час побудови параметричних моделей.

Підсумок етапу підготовки даних. У межах підготовки вибірки виконано імпутацію пропущених значень методом kNN, формування бінарних текстових ознак на основі описів оголошень, геокодування адрес із розрахунком відстані до центру міста, вінзоризацію цільових змінних, нормалізацію числових показників та кореляційний аналіз.

У результаті сформовано аналітичний набір даних, придатний для подальшого застосування параметричних та ансамблевих методів машинного навчання з метою оцінювання важливості ознак і моделювання ціноутворення на вторинному ринку житлової нерухомості Києва.

Формалізація задачі прогнозування

Задачу прогнозування вартості житлової нерухомості формалізуємо як задачу багатовимірної регресії [14]. Нехай $X = (x_1, \ldots, x_n)$ – вектор ознак, що включає фізичні, локаційні та інфраструктурні характеристики об'єкта, а також предиктори, отримані внаслідок NLP-аналізу текстових описів. Цільова змінна y розглядається у двох варіаціях: загальна ціна *price* та ціна за квадратний метр *price_sm*.

Модель має вигляд:

$$y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

$f(\cdot)$ – невідома функціональна залежність, що апроксимується параметричною (лінійна регресія) або алгоритмічною моделлю машинного навчання, ε_i – випадкова похибка.

У випадку лінійної регресії:

$$f(X_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{2}$$

що відповідає класичній постановці регресійної моделі [14]. Для ансамблевих алгоритмів і нейронних мереж функція формується як нелінійна композиція базових моделей.

Для забезпечення коректного оцінювання якості прогнозних моделей сформований набір даних було поділено на тренувальну та тестову вибірки у пропорції 80 % та 20 % відповідно. До тренувальної вибірки увійшло 11 197 спостережень, до тестової – 2 800 спостережень.

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

Навчання моделей здійснювалося виключно на тренувальній вибірці, тоді як оцінювання показників точності (R^2, RMSE, MAE, MAPE) виконувалося на тестовій вибірці, яка не використовувалася під час навчання. Такий підхід дозволяє отримати більш об'єктивну оцінку узагальнювальної здатності моделей та мінімізувати ризик перенавчання.

Базова параметрична модель (лінійна регресія)

Лінійна регресія. Модель загальної ціни (price). Оцінювання впливу пояснювальних змінних на загальну вартість житлової нерухомості здійснено за допомогою множинної лінійної регресії. Відбір змінних проводився покроковою процедурою з мінімізацією інформаційного критерію

Попередньо виконано діагностику мультиколінеарності із застосуванням коефіцієнта інфляції дисперсії (VIF). На початковому етапі максимальне значення VIF становило 6,14. Після виключення змінної *area_living* повторна оцінка показала зниження максимального значення VIF до 4,08, що не вказує на наявність критичної мультиколінеарності.

Загальний вигляд моделі:

$$price_i = \beta_0 + \sum_{j=1}^{18} \beta_j x_{ij}, \tag{3}$$

де $price_i$ – загальна ціна і-го об'єкта (тис. грн), x_{ij} – пояснювальні змінні, ε_i – випадкова похибка.

Із 30 сформованих ознак до підсумкової специфікації регресійної моделі увійшли 18 предикторів, відібраних за процедурою покрокової регресії (stepwise selection) для оптимізації прогнозної здатності. Створена модель включає ключові фактори: загальна площа (area_total), площа житлових кімнат і кухні (area_living, area_kitchen), кількість кімнат, поверховість (stock, stock_total), вік будинку, матеріал стін (wall), відстань до центру (distance_to_center) та інфраструктурні ознаки (balcony, elevator, parking, school_nearby, park_nearby, appliances, transport_nearby). Оцінена модель має вигляд (4):

$$\begin{aligned} price = {} & 0{,}14 - 0{,}15 \cdot rooms + 0{,}84 \cdot area\_total + 0{,}06 \cdot area\_kitchen - 0{,}01 \cdot stock - \\ & - 0{,}02 \cdot stock\_total + 0{,}02 \cdot house\_age - 0{,}05 \cdot wall - 0{,}15 \cdot distance\_to\_center - \\ & - 0{,}01 \cdot project + 0{,}01 \cdot bad\_proposal - 0{,}02 \cdot balcony + 0{,}01 \cdot \text{elevator} - 0{,}02 \cdot \text{parking} - \\ & - 0{,}03 \cdot school\_nearby + 0{,}01 \cdot park\_nearby + 0{,}04 \cdot appliances - 0{,}03 \cdot transport\_nearby + \varepsilon. \end{aligned} \tag{4}$$

Бінарні змінні приймають значення 1 – наявність ознаки, 0 – її відсутність. параметри множинної лінійної регресії, їх стандартні похибки та короткий опис наведено в табл. 2.

Результати оцінювання моделі свідчать, що найбільший позитивний коефіцієнт серед безперервних змінних має area_total, що узгоджується з результатами кореляційного аналізу. Змінна distance_to_center демонструє статистично значущий від'ємний ефект, що відображає просторову залежність вартості житла від віддаленості від центральної частини міста.

Інфраструктурні характеристики мають різноспрямовані коефіцієнти. Зокрема, змінна appliances характеризується позитивною оцінкою параметра, тоді як transport_nearby – від'ємною. Кількість кімнат (rooms) має від'ємний коефіцієнт за фіксованої загальної площі, що пов'язано зі структурою моделі та взаємозв'язком між площею і плануванням об'єкта.

Перед обчисленням показників точності прогнозу (R^2, RMSE, MAE, MAPE) усі прогнозні значення було денормалізовано з метою забезпечення коректності інтерпретації метрик, чутливих до масштабу даних.

Модель пояснює 73,5 % варіації цін ($R^2 = 0{,}735$), що свідчить про достатню пояснювальну здатність специфікації. Водночас тест Бройша–Пагана ($BP = 2344{,}5$; $p < 2{,}2 \cdot 10^{-16}$) виявив наявність гетероскедастичності залишків. Тест Дарбіна–Уотсона ($DW = 1{,}92$; $p = 1{,}886 \cdot 10^{-6}$) засвідчив статистично значущу, але помірну автокореляцію.

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

Таблиця 2

Оцінки параметрів множинної лінійної регресії (модель price)

Змінна	β	SE	Опис
(Intercept)	0,1365***	0,0070	Константа моделі
rooms	–0,1499***	0,0061	Кількість кімнат
area_total	0,8436***	0,0087	Загальна площа, м^2
area_kitchen	0,0566***	0,0056	Площа кухні, м^2
stock	–0,0119*	0,0049	Поверх розташування
stock_total	–0,0191***	0,0051	Загальна поверховість будинку
house_age	0,0238***	0,0059	Вік будинку, років
wall	–0,0502***	0,0041	Тип матеріалу стін (числове кодування)
distance_to_center	–0,1461***	0,0039	Відстань до центру міста, км
project	–0,0137*	0,0056	Тип/серія будинку (числове кодування)
bad_proposal	0,0109**	0,0038	Індикатор проблемної пропозиції (1 – так, 0 – ні)
balcony	–0,0215***	0,0026	Наявність балкона (1 – так, 0 – ні)
elevator	0,0076**	0,0029	Наявність ліфта (1 – так, 0 – ні)
parking	–0,0161***	0,0033	Наявність паркування (1 – так, 0 – ні)
school_nearby	–0,0279***	0,0025	Наявність школи поблизу (1 – так, 0 – ні)
park_nearby	0,0128***	0,0026	Наявність парку поблизу (1 – так, 0 – ні)
appliances	0,0359***	0,0050	Наявність побутової техніки (1 – так, 0 – ні)
transport_nearby	–0,0302***	0,0024	Наявність транспортної доступності (1 – так, 0 – ні)

Примітка. β – оцінка параметра моделі (метод найменших квадратів);
SE – стандартна похибка оцінки;
** $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$.*

Показники точності прогнозу становлять: RMSE = 8 257 650 грн, MAE = 6 640 116 грн, MAPE = 39,05 %. Отримані значення відображають середню абсолютну та відносну похибку моделі в межах досліджуваної вибірки.

Модель ціни за квадратний метр (price_sm). Моделювання питомої вартості квадратного метра (price_sm) здійснено із застосуванням тієї ж процедури оцінювання – множинної лінійної регресії з покроковим відбором змінних за критерієм AIC та попередньою перевіркою мультиколінеарності (VIF).

Загальний вигляд моделі відповідає формулі (3), де залежною змінною виступає *price_sm* – варстість квадратного метра житла у тис. грн.

$$\begin{aligned} price\_sm = {} & 0.46 - 0.20 \cdot rooms + 0.35 \cdot area\_total + 0.13 \cdot area\_kitchen - 0.03 \cdot \text{stock} - \\ & - 0.01 \cdot stock\_total - 0.08 \cdot wall - 0.27 \cdot distance\_to\_center - 0.02 \cdot project + \\ & + 0.03 \cdot bad\_proposal - 0.03 \cdot balcony + 0.01 \cdot elevator - 0.02 \cdot parking + \\ & + 0.02 \cdot security - 0.06 \cdot school\_nearby + 0.03 \cdot park\_nearby + 0.08 \cdot appliances - \\ & - 0.05 \cdot transport\_nearby + \varepsilon. \end{aligned} \quad (5)$$

Результати оцінювання моделі *price_sm* показують, що найбільший позитивний коефіцієнт серед безперервних змінних має *area_total*. Змінна *distance_to_center* характеризується статистично значущим від'ємним коефіцієнтом (≈ –0,27), що відображає зниження питомої вартості зі збільшенням віддаленості від центральної частини міста.

Інфраструктурні характеристики демонструють різноспрямовані оцінки параметрів. Позитивні коефіцієнти отримано для змінних *appliances* та *security*, тоді як *transport_nearby*, *balcony* і *parking* мають від'ємні оцінки. Такі результати узгоджуються зі структурою вибірки та специфікацією моделі.

Коефіцієнт детермінації становить R^2 = 0,384, що свідчить про помірну пояснювальну здатність моделі. Тест Бройша–Пагана (BP = 1801,3; $p < 0,001$) вказує на наявність

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

гетероскедастичності залишків. Тест Дарбіна–Уотсона (DW = 1,877; $p < 0{,}001$) засвідчує статистично значущу, але помірну автокореляцію.

Показники точності прогнозу становлять: RMSE = 29 703, MAE = 22 745, MAPE = 32,56 %. Отримані значення характеризують середню абсолютну та відносну похибку моделі у межах досліджуваної вибірки.

Порівняння моделі загальної ціни та моделі ціни за квадратний метр свідчить про різний рівень пояснювальної здатності та наявність структурних особливостей у даних. Це обґрунтовує доцільність застосування більш гнучких регресійних підходів і робастних процедур оцінювання на наступному етапі аналізу.

Ансамблеві алгоритми градієнтного бустингу

Для моделювання вартості житлової нерухомості застосовано алгоритм градієнтного бустингу дерев рішень XGBoost, що дозволяє враховувати нелінійні залежності та взаємодії між ознаками. Важливість змінних оцінювалася за метрикою Gain, яка відображає середнє зменшення функції втрат при використанні відповідної ознаки для розбиття вузлів у процесі побудови ансамблю дерев. Оскільки у моделі використано функцію втрат, засновану на середньоквадратичній помилці, значення Gain характеризує внесок змінної у зниження похибки прогнозування.

Модель загальної ціни (price). За результатами оцінювання найбільше значення Gain має змінна *area_total*. Високий внесок також демонструють *distance_to_center*, *complex* та *house_age*. Інші характеристики – площа кухні, поверховість, матеріал стін та інфраструктурні ознаки – мають менший, але статистично помітний внесок у структуру прогнозу.

Модель характеризується коефіцієнтом детермінації R^2 = 0,874 та відносною похибкою MAPE = 17,8 %. Значення RMSE становить 1 995,44 тис. грн, MAE – 1 188,1 тис. грн. Отримані показники свідчать про істотне підвищення точності прогнозування порівняно з лінійною регресією.

Результати важливості ознак наведено на рис. 4.

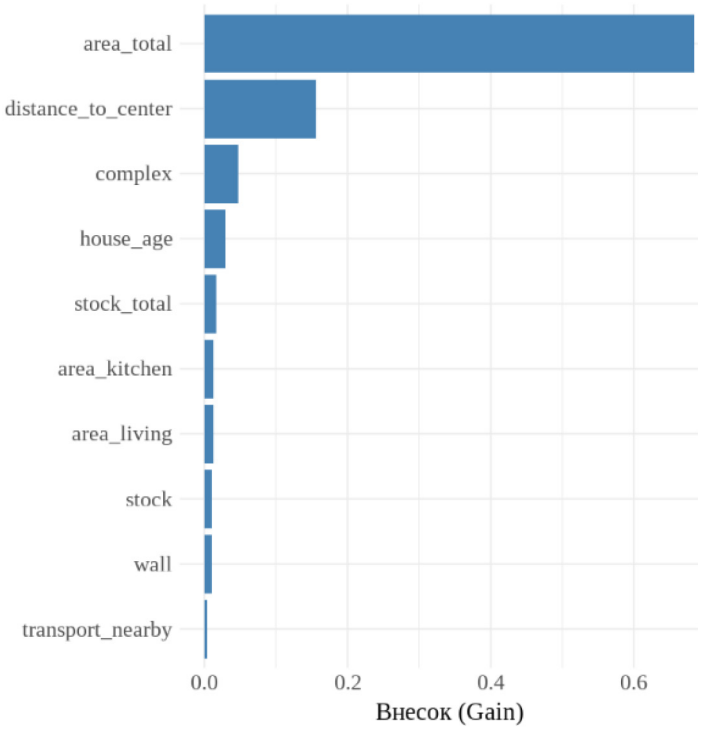

Рис. 4. Графік важливості ознак XGBoost для ціни нерухомості

Джерело: створено авторами.

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

Модель ціни за квадратний метр (price_sm). Для моделювання питомої вартості житла найбільший внесок у прогноз за метрикою Gain мають змінні distance_to_center, complex, house_age та area_total. Це свідчить про вагомість локаційних характеристик і параметрів будівлі навіть при оцінюванні ціни за квадратний метр.

Додатковий вплив демонструють wall, area_kitchen, area_living, stock_total, а також характеристики проєкту та планування. Інфраструктурні змінні – renovation_type, appliances, balcony, elevator, parking та показники навколишнього середовища – мають менші, але стабільні значення важливості. Таким чином, структура прогнозу формується поєднанням просторових, конструктивних і додаткових характеристик об'єкта.

Модель характеризується коефіцієнтом детермінації $R^2 = 0{,}756$, що означає пояснення 75,6 % варіації ціни за м² у вибірці. Відносна похибка прогнозу становить MAPE = 17,8 %. Значення RMSE дорівнює 18,63 тис. грн/м², MAE – 12,96 тис. грн/м². Отримані показники підтверджують покращення точності порівняно з лінійною специфікацією.

Результати XGBoost свідчать про те, що найбільший внесок у формування прогнозу мають локаційні та базові фізичні характеристики об'єкта. Інфраструктурні змінні, комплектація та планування мають менші значення важливості та виконують уточнювальну функцію в межах моделі. Загалом найбільш впливовими залишаються площа об'єкта, відстань до центру, належність до житлового комплексу та характеристики будівлі, що узгоджується зі структурою отриманих оцінок важливості.

Для перевірки стійкості результатів і порівняння з XGBoost додатково застосовано алгоритм LightGBM – реалізацію градієнтного бустингу дерев рішень, оптимізовану для роботи з табличними даними. Важливість змінних визначалась на основі їхнього сумарного внеску у зменшення функції втрат під час побудови ансамблю дерев. Оскільки у задачі регресії використовується функція втрат, пов'язана із середньоквадратичною помилкою, показник важливості відображає агреговане зниження похибки прогнозування, яке забезпечує відповідна змінна.

Таким чином, значення важливості у LightGBM інтерпретується як відносний внесок ознаки у покращення якості прогнозу в межах моделі.

Модель загальної ціни (price). За результатами оцінювання найбільше значення важливості має змінна area_total. Суттєвий внесок також демонструє distance_to_center. Високі показники важливості мають complex, area_living, hist_district, house_age, admin_district та area_kitchen. Змінні stock_total і stock мають менший, але стабільний внесок у структуру прогнозу.

Коефіцієнт детермінації становить $R^2 = 0{,}888$, що перевищує відповідний показник для XGBoost. Відносна похибка прогнозу дорівнює MAPE ≈ 17,1 %. Значення RMSE становить 1 884,27 тис. грн, MAE – 1 094,35 тис. грн. Отримані показники свідчать про високу узгодженість прогнозованих та фактичних значень у межах вибірки.

Модель ціни за квадратний метр (price_sm). За результатами LightGBM найбільше значення важливості для price_sm має змінна distance_to_center, що вказує на вагомість локаційного фактора у структурі прогнозу питомої вартості. Високі показники важливості також демонструють hist_district, complex, house_age та конструктивні характеристики будівлі.

Коефіцієнт детермінації становить $R^2 = 0{,}793$, тобто модель пояснює 79,3 % варіації ціни за м² у вибірці. Значення RMSE дорівнює 11,65, MAE – 17,14, а MAPE – 15,7 %. Отримані показники свідчать про підвищення точності порівняно з лінійною специфікацією та узгоджуються з результатами інших ансамблевих моделей.

Результати важливості ознак наведено на рис. 6.

Отримані результати LightGBM загалом узгоджуються з оцінками, отриманими за допомогою XGBoost. В обох випадках найбільші значення важливості мають площа об'єкта та локаційні характеристики, тоді як параметри будівлі та інфраструктурні змінні мають менший, але стабільний внесок у структуру прогнозу.

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

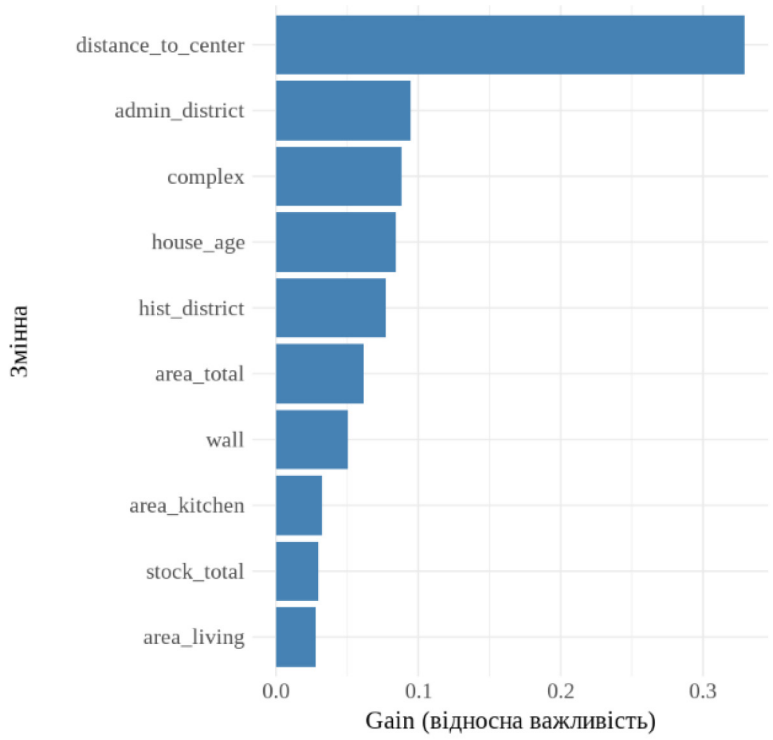

Рис. 5. Графік важливості змінних моделі LightGBM ціни за метр квадратний нерухомості

Джерело: створено авторами.

Ансамблевий метод Random Forest

Для додаткової перевірки стійкості результатів застосовано ансамблевий алгоритм Random Forest, який формує прогноз шляхом усереднення великої кількості дерев рішень. Такий підхід зменшує варіативність оцінок та дозволяє враховувати нелінійні залежності між ознаками.

Важливість змінних оцінювалась за показником IncNodePurity (Increase in Node Purity), що відображає сумарне зменшення неоднорідності вузлів у процесі розбиття за відповідною змінною. У задачі регресії цей показник базується на зменшенні суми квадратів залишків (RSS) під час побудови дерев. Більше зниження помилки відповідає більшому внеску змінної у формування прогнозу.

Модель загальної ціни (price). Найбільші значення важливості отримано для змінної area_total. Суттєвий внесок також демонструють area_living, admin_district та distance_to_center. Помітний вплив мають rooms, area_kitchen, complex та house_age.

Коефіцієнт детермінації становить $R^2 = 0{,}878$. Значення RMSE дорівнює 1 159,92, MAE – 1 961,38, а MAPE – 18,5 %. Отримані показники є співставними з результатами інших ансамблевих алгоритмів.

Модель ціни за квадратний метр (price_sm). Для price_sm найбільше значення важливості має змінна admin_district. Високий внесок також демонструють distance_to_center, complex та house_age. Додатковий вплив мають площинні характеристики квартири та конструктивні параметри будівлі.

Коефіцієнт детермінації становить $R^2 = 0{,}786$. Значення RMSE дорівнює 11,98, MAE – 17,43, MAPE – 16,4 %. Результати узгоджуються з оцінками, отриманими для інших ансамблевих моделей.

Загалом Random Forest підтверджує структуру важливості ознак, виявлену XGBoost і LightGBM: найбільший внесок мають площинні та локаційні характеристики, тоді як додаткові параметри уточнюють прогноз у межах моделі.

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

Штучна нейронна мережа

Для врахування можливих нелінійних залежностей між характеристиками житла та його вартістю застосовано штучну нейронну мережу типу багатошарового перцептрону. Така архітектура дозволяє моделювати складні взаємозв'язки між змінними та формувати узагальнений прогноз цільового показника.

Для оцінювання відносної важливості предикторів використано алгоритм Гарсона (Garson's Algorithm), реалізований у пакеті NeuralNetTools. Підхід базується на аналізі вагових коефіцієнтів між вхідним, прихованим та вихідним шарами мережі. Відносна важливість змінної визначається шляхом декомпозиції абсолютних значень ваг зв'язків і агрегування їх внеску через приховані нейрони до вихідного шару. Отриманий показник (relative importance) нормується таким чином, що сума важливостей усіх предикторів дорівнює 1 (або 100 %), що дозволяє порівнювати їх внесок у межах моделі.

Модель загальної ціни (price). За результатами моделювання найбільші значення відносної важливості мають змінні admin_district, hist_district, distance_to_center, area_total, house_age та complex. Помітний внесок також демонструють wall, rooms, stock_total та elevator. Отримана структура важливості відображає поєднання локаційних та конструктивних характеристик об'єкта.

Коефіцієнт детермінації становить R^2 = 0,835. Середня абсолютна помилка дорівнює 2 263,17, середнє квадратичне відхилення – 1 468,82, а відносна помилка MAPE – близько 25,7 %. Отримані показники свідчать про здатність моделі відтворювати загальні закономірності варіації ціни у вибірці, хоча точність є нижчою порівняно з ансамблевими алгоритмами.

Модель ціни за квадратний метр (price_sm). Для price_sm найбільше значення відносної важливості має змінна complex. Високі показники також отримано для distance_to_center, house_age, admin_district та hist_district. Меншу відносну важливість мають area_total, stock_total, wall, project та rooms, що вказує на диференційований внесок просторових і конструктивних характеристик у формування питомої вартості.

Коефіцієнт детермінації становить R^2 = 0,637. Значення RMSE дорівнює 16,07, MAE – 22,47, а MAPE – близько 22,5 %. Порівняно з бустинговими моделями точність є нижчою, що відображає особливості налаштування мережі та структуру доступних даних.

Загалом результати нейронної мережі узгоджуються з оцінками, отриманими за допомогою Random Forest, XGBoost та LightGBM: найбільший внесок у прогноз мають локаційні характеристики, параметри будівлі та базові площинні змінні. Додаткові індикатори, сформовані на основі текстових описів і закодовані у вигляді бінарних ознак, уточнюють структуру прогнозу в межах моделі.

Порівняльний аналіз моделей

У межах дослідження виконано порівняння лінійної регресії, ансамблевих алгоритмів (XGBoost, LightGBM, Random Forest) та штучної нейронної мережі для задач прогнозування загальної ціни (price) та ціни за квадратний метр (price_sm). Оцінювання здійснювалося за показниками R^2, RMSE, MAE та MAPE.

Найвищі значення R^2 для обох цільових змінних продемонструвала модель LightGBM. Для загальної ціни R^2 становить 0,888, MAPE – 17,1 %, що є найменшою відносною похибкою серед розглянутих моделей. Для ціни за квадратний метр LightGBM також показала найкращі результати (R^2 = 0,793; MAPE = 15,7 %).

XGBoost та Random Forest продемонстрували співставні показники точності. Для змінної price їхні значення R^2 становлять 0,874 та 0,878 відповідно, а MAPE – 19,2 % і 18,5 %. Для price_sm значення R^2 перебувають у межах 0,756–0,786, а MAPE – 16,4–17,8 %. Отримані результати свідчать про стабільність ансамблевих алгоритмів при моделюванні ринку нерухомості.

Штучна нейронна мережа продемонструвала нижчі показники точності порівняно з бустинговими моделями (R^2 = 0,835 для price та 0,637 для price_sm), що може бути пов'язано з особливостями структури даних та параметризації мережі.

https://doi.org/10.32782/mathematical-modelling/2026-9-1-15 ISSN 2618-0332 (print), 2618-0340 (online)

Лінійна регресія показала найменші значення R^2, особливо для моделі price_sm ($R^2 = 0{,}384$), що вказує на обмеження параметричної специфікації при моделюванні нелінійних залежностей.

Результати оцінювання моделей наведено в табл. 3.

Таблиця 3

Результати оцінювання моделей прогнозування загальної ціни та ціни нерухомості за м²

Модель	RMSE	R^2	MAE	MAPE (%)
Linear Regression (price)	**8 257 650**	**0.735**	**6 640 116**	**39.0**
Linear Regression (price_sm)	**29 703**	**0.384**	**22 745.4**	**32.6**
XGBoost (price)	1 995.44	0.874	1 188.1	19.2
XGBoost (price_sm)	18.6275	0.756	12.9559	17.8
LightGBM (price)	**1 884.27**	**0.888**	**1 094.35**	**17.1**
LightGBM (price_sm)	**17.139**	**0.793**	**11.6511**	**15.7**
Random Forest (price)	1 961.38	0.878	1 159.92	18.5
Random Forest (price_sm)	17.4349	0.786	11.9831	16.4
Neural Network (price)	2 263.17	0.835	1 468.82	25.7
Neural Network (price_sm)	22.4717	0.637	16.0666	22.5

Джерело: створено авторами.

Висновки

У результаті дослідження встановлено, що ансамблеві алгоритми градієнтного бустингу забезпечують найвищу точність прогнозування вартості житлової нерухомості на вторинному ринку Києва. Найкращі результати отримано для моделі LightGBM як для загальної ціни, так і для ціни за квадратний метр.

Аналіз важливості ознак показав, що для загальної вартості об'єкта найбільший внесок мають площинні характеристики (area_total), локаційні змінні (distance_to_center, admin_district) та параметри будівлі (house_age, complex). Для питомої вартості більш вагомими є локаційні характеристики та адміністративна прив'язка, тоді як площинні параметри та конструктивні особливості мають уточнювальний характер.

Ознаки, сформовані на основі текстових описів і закодовані у вигляді бінарних змінних, демонструють додатковий внесок у структуру прогнозу, що підтверджує доцільність розширення набору предикторів на етапі інженерії ознак.

Отримані результати свідчать про доцільність використання ансамблевих алгоритмів для задач прогнозування вартості житла та подальшого аналізу структури факторів ціноутворення на локальних ринках нерухомості.

Список використаної літератури

1. Національний банк України. Звіт про фінансову стабільність. Грудень 2025 року [Електронний ресурс]. Київ : Національний банк України, 2025. URL: https://bank.gov.ua/admin_uploads/article/FSR_2025-H2.pdf (дата звернення: 27.01.2026).
2. Деякі питання забезпечення функціонування Єдиної державної електронної системи у сфері будівництва : Постанова Кабінету Міністрів України від 23.06.2021 № 681 : станом на 30.12.2025 [Електронний ресурс] // База даних «Законодавство України». URL: https://zakon.rada.gov.ua/go/681-2021-%D0%BF (дата звернення: 27.01.2026).
3. Про оцінку майна, майнових прав та професійну оціночну діяльність в Україні : Закон України від 12.07.2001 № 2658-III [Електронний ресурс] // База даних «Законодавство України». URL: https://zakon.rada.gov.ua/go/2658-14 (дата звернення: 27.01.2026).
4. Пашкевич О., Ващищак С., Бойчук А., Стисло Т., Демчина М. Застосування моделей машинного навчання для прогнозування цін на ринку нерухомості. *Вісник Хмельницького*

національного університету. Серія : Технічні науки. 2022. № 5 (313). С. 265–273. DOI: https://doi.org/10.31891/2307-5732-2022-313-5-265-273
5. Hoxha V., Shala A. Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina's real estate market. *International Journal of Housing Markets and Analysis*. 2025. Vol. 18, No. 3. P. 694–711. DOI: https://doi.org/10.1108/IJHMA-09-2023-0120
6. Moreno-Foronda I., Sánchez-Martínez M.-T., Pareja-Eastaway M. Comparative Analysis of Advanced Models for Predicting Housing Prices: A Review. *Urban Science*. 2025. Vol. 9, No. 2. Art. 32. DOI: https://doi.org/10.3390/urbansci9020032
7. Alkan T., Dokuz Y., Ecemiş A., Bozdağ A., Durduran S. S. Using machine learning algorithms for predicting real estate values in tourism centers. *Soft Computing*. 2023. Vol. 27. P. 2601–2613. DOI: https://doi.org/10.1007/s00500-022-07579-7
8. Yazdani M. Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction [Електронний ресурс] // arXiv. 2021. DOI: https://doi.org/10.48550/arXiv.2110.07151
9. Jha S. B., Babiceanu R. F., Pandey V., Jha R. K. Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study [Електронний ресурс] // arXiv. 2020. DOI: https://doi.org/10.48550/arXiv.2006.10092
10. Верес О., Шимоняк А. Прогнозування вартості нерухомості з використанням засобів машинного навчання. *Information Systems and Networks*. 2024. Вип. 15. С. 140–158. DOI: https://doi.org/10.23939/sisn2024.15.140
11. Zhang H., Li Y., Branco P. Describe the house and I will tell you the price: House price prediction with textual description data. *Natural Language Engineering*. 2024. Vol. 30, Iss. 4. P. 661–695. DOI: https://doi.org/10.1017/S1351324923000360
12. R Core Team. R: A language and environment for statistical computing [Електронний ресурс]. Vienna : R Foundation for Statistical Computing, 2024. URL: https://www.R-project.org/ (дата звернення: 27.01.2026).
13. Kyrychenko R. Kyiv secondary (resale) residential real estate [Електронний ресурс] : dataset. Zenodo, 2026. DOI: https://doi.org/10.5281/zenodo.18413277
14. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: With Applications in R. 2nd ed. New York : Springer, 2021. 607 p. DOI: https://doi.org/10.1007/978-1-0716-1418-1

References

1. National Bank of Ukraine. (2025). Zvit pro finansovu stabilnist. Hruden 2025 roku [Financial stability report. December 2025]. https://bank.gov.ua/admin_uploads/article/FSR_2025-H2.pdf [in Ukrainian].
2. Cabinet of Ministers of Ukraine. (2021, June 23). Deiaki pytannia zabezpechennia funktsionuvannia Yedynoi derzhavnoi elektronnoi systemy u sferi budivnytstva [Some issues of ensuring the functioning of the Unified State Electronic System in the field of construction] (Postanova No. 681, stanom na 30.12.2025). https://zakon.rada.gov.ua/go/681-2021-%D0%BF [in Ukrainian].
3. Verkhovna Rada of Ukraine. (2001, July 12). Pro otsinku maina, mainovykh prav ta profesiinu otsinochnu diialnist v Ukraini [On the valuation of property, proprietary rights and professional valuation activity in Ukraine] (Zakon Ukrainy No. 2658-III). https://zakon.rada.gov.ua/go/2658-14 [in Ukrainian].
4. Pashkevych, O., Vashchyshchak, S., Boichuk, A., Styslo, T., & Demchyna, M. (2022). Zastosuvannia modelei mashynnoho navchannia dlia prohnozuvannia tsin na rynku nerukhomosti [Application of machine learning models for predicting real estate prices]. *Visnyk Khmelnytskoho natsionalnoho universytetu. Seriia: Tekhnichni nauky*, *313*(5), 265–273. https://doi.org/10.31891/2307-5732-2022-313-5-265-273 [in Ukrainian].

5. Hoxha, V., & Shala, A. (2025). Comparative analysis of machine learning models in predicting housing prices: A case study of Prishtina's real estate market. *International Journal of Housing Markets and Analysis*, *18*(3), 694–711. https://doi.org/10.1108/IJHMA-09-2023-0120 [in English].
6. Moreno-Foronda, I., Sánchez-Martínez, M.-T., & Pareja-Eastaway, M. (2025). Comparative analysis of advanced models for predicting housing prices: A review. *Urban Science*, *9*(2), Article 32. https://doi.org/10.3390/urbansci9020032 [in English].
7. Alkan, T., Dokuz, Y., Ecemiş, A., Bozdağ, A., & Durduran, S. S. (2023). Using machine learning algorithms for predicting real estate values in tourism centers. *Soft Computing*, *27*, 2601–2613. https://doi.org/10.1007/s00500-022-07579-7 [in English].
8. Yazdani, M. (2021). Machine learning, deep learning, and hedonic methods for real estate price prediction. arXiv. https://doi.org/10.48550/arXiv.2110.07151 [in English].
9. Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). Housing market prediction problem using different machine learning algorithms: A case study. arXiv. https://doi.org/10.48550/arXiv.2006.10092 [in English].
10. Veres, O., & Shymoniak, A. (2024). Prohnozuvannia vartosti nerukhomosti z vykorystanniam zasobiv mashynnoho navchannia [Forecasting real estate value using machine learning tools]. *Information Systems and Networks*, 15, 140–158. https://doi.org/10.23939/sisn2024.15.140 [in Ukrainian].
11. Zhang, H., Li, Y., & Branco, P. (2024). Describe the house and I will tell you the price: House price prediction with textual description data. *Natural Language Engineering*, 30(4), 661–695. https://doi.org/10.1017/S1351324923000360. [in English].
12. R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/ [in English].
13. Kyrychenko, R. (2026). Kyiv secondary (resale) residential real estate [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18413277 [in English].
14. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1 [in English].

Клебан Юрій Вікторович – старший викладач кафедри інформаційних технологій та аналітики даних Національного університету «Острозька академія». E-mail: yuriy.kleban@oa.edu.ua, ORCID: 0000-0002-7070-5175.

Лепська Ліна Олександрівна – студентка кафедри інформаційних технологій та аналітики даних Національного університету «Острозька академія». E-mail: lina.lepska@oa.edu.ua, ORCID: 0009-0008-1143-1769.

Kleban Yurii Viktorovych – Senior Lecturer at the Department of Information Technologies and Data Analytics of the National University of Ostroh Academy. E-mail: yuriy.kleban@oa.edu.ua, ORCID: 0000-0002-7070-5175.

Lepska Lina Oleksandrivna – Student at the Department of Information Technologies and Data Analytics of the National University of Ostroh Academy. E-mail: lina.lepska@oa.edu.ua, ORCID: 0009-0008-1143-1769.

Дата першого надходження статті до видання: 05.03.2026
Дата прийняття статті до друку після рецензування: 23.04.2026
Дата публікації (оприлюднення) статті: 01.07.2026