УДК 004.415.2
DOI https://doi.org/10.32782/2663-5941/2023.6/18

***Олещенко Л.М.***
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

***Трушина Д.В.***
Таврійський національний університет імені В.І. Вернадського

# ПРОГРАМНИЙ МЕТОД ПРОГНОЗУВАННЯ ВАРТОСТІ НЕРУХОМОСТІ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ ТА РЕГРЕСІЙНОГО АНАЛІЗУ ДАНИХ

*У статті розглядаються особливості програмної реалізації регресійної моделі для прогнозування вартості нерухомості з використанням методів машинного навчання. Для вибору моделі прогнозування цін на житло проведено аналіз публікацій та існуючих програмних рішень для прогнозування вартості нерухомості. У даному дослідженні для практичної реалізації запропонованого програмного методу використано мову програмування Python, бібліотеки Pandas, Matplotlib, Seaborn, Scikit-learn, та NumPy. Для дослідження побудови моделі регресії було обрано датасет «Linear Regression – House Price Predictions», обсягом 514 КВ, що містить 4601 запис. Здійснено навчання моделі, оцінку її точності та порівняння з іншими методами прогнозування з використанням технологій машинного навчання.*

*У статті наведено дослідження областей застосування моделі для прогнозування вартості нерухомості на основі регресійного аналізу даних, формування вимог до програмного забезпечення, вибір мови програмування, розробка структури та навчання алгоритму. Розглянуті технологічні підходи включають обробку та підготовку даних, вибір функцій, регресійне моделювання, навчання та оцінку моделі, інтеграцію штучного інтелекту та машинного навчання, валідацію та оптимізацію моделі.*

*У результаті дослідження створено програмний метод для прогнозування вартості нерухомості на основі регресійного аналізу даних, який надає графічне відображення переліку факторів з кількісною оцінкою їх впливу на вартість нерухомості. Переваги цього дослідження є важливими з точки зору програмної реалізації методу регресійного аналізу даних та можливості майбутнього використання запропонованого методу для прогнозування ринку нерухомості в Україні. Розробивши модель на основі регресійного аналізу даних, фахівці з нерухомості отримають надійний інструмент для прогнозування вартості нерухомості різного типу. Це дозволить їм приймати зважені рішення щодо інвестицій, цінових стратегій і проектів розвитку. Крім того, власники нерухомості краще розумітимуть фактори, що впливають на вартість їхньої нерухомості, що дозволить приймати обґрунтовані рішення щодо покращення житлових умов.*

***Ключові слова:*** *технології програмування систем штучного інтелекту, мова програмування Python, прогнозування, машинне навчання, ринок нерухомості, великі дані, оброблення даних, регресійний аналіз даних.*

**Постановка проблеми.** У сучасному світі нерухомість визначається не лише фізичними будівлями та землею, а й даними та аналітикою. З розвитком технологій аналізу даних і машинного навчання з'явилася можливість покращити прогнозування вартості нерухомості на основі широкого спектру факторів. У прогнозуванні вартості нерухомості наявні такі проблеми, як нестабільність ринку, багатовимірність даних, проблеми з недостатньою інформацією, точність та надійність даних, а також регуляторні аспекти. Ринок нерухомості може бути вкрай нестабільним і піддаватися змінам через різноманітні фактори, такі як економічна кон'юнктура, політичні події, зміни в законодавстві. Це ускладнює точне прогнозування вартості нерухомості. Дані про нерухомість можуть бути багатовимірними, з великою кількістю впливових факторів (земельні площі, розташування, інфраструктура тощо). Відбір найбільш важливих змінних для включення в модель є складною задачею. Деякі аспекти нерухомості досить важко або навіть неможливо виміряти чи кількісно визначити. Наприклад, атмосфера в районі, екологічна ситуація, комфортність для проживання – ці аспекти можуть впливати на ціну, їх важко врахувати в аналізі. Доступні дані можуть

бути неповними або неправильними, що може призвести до неточностей у прогнозній моделі. Важливо також враховувати якість та достовірність джерел даних. Зміни в законодавстві чи податковій політиці можуть значно вплинути на ринок нерухомості. Прогнозування вартості повинно враховувати різні регуляторні аспекти. Актуальність дослідження полягає в тому, що точне прогнозування вартості нерухомості має важливе значення для різних суб'єктів ринку: інвесторів, забудовників, споживачів, страхових компаній та інших учасників. Здатність ефективно прогнозувати ціни може сприяти прийняттю обґрунтованих рішень у галузі нерухомості, що має важливий вплив на економіку та життя людей.

Для досягнення поставленої мети необхідне виконання наступних завдань: аналіз змін цін на ринку нерухомості протягом певного періоду; вивчення впливу різних факторів, таких як площа, кількість кімнат, розташування та інфраструктура тощо на вартість нерухомості; визначення та аналіз основних тенденцій у розвитку ринку нерухомості; розробка моделі для прогнозування майбутніх цін на нерухомість на основі історичних даних; визначення та оцінка різних факторів (економісні, соціокультурні, політичні), які можуть впливати на ціни на нерухомість; оцінка точності та ефективності розробленої моделі прогнозування.

**Метою статті** є розробка програмного методу прогнозування вартості нерухомості на основі регресійного аналізу даних та сучасних технологій машинного навчання.

**Аналіз публікацій та існуючих програмних рішень для прогнозування вартості нерухомості.** Автори статті [1] використовують підхід часових рядів для аналізу факторів, що впливають на індекс реальних цін на житло в США, і застосовують комбінацію моделей оптимальності для прогнозування індексу. У роботі [2] представлено порівняльний аналіз динамічних факторних моделей та моделей LBVAR для прогнозування цін на житло. Модель динамічних факторів DFM передбачає, що ціни на житло визначаються змінами фундаментальних факторів і ринковими рухами з часом. З іншого боку, модель LBVAR включає в себе комбінацію моделей VAR Ласкіна-Бізекера (LBVAR) і моделей векторної авторегресії (VAR), щоб зафіксувати динамічний тренд і короткострокову волатильність цін на житло. Автори порівняли показники обох моделей, використовуючи дані з США за період 1995–2010 років. Стаття [3] описує дослідження між пошуковою активністю в мережі Інтернет та змінами цін на житлову нерухомість. У дослідженні використовувалися дані Google Trends для вимірювання пошукової активності та відстежувалися зміни цін на житлову нерухомість в Остіні, штат Техас. Результати показують, що існує позитивна кореляція між пошуковою активністю в мережі Інтернет та цінами на житлову нерухомість. Це вказує на те, що пошукову активність в Інтернеті можна використовувати як прогноз майбутніх змін цін на ринку житлової нерухомості. У статті [4] результати аналізу підтвердили використання штучних нейронних мереж, опорної векторної регресії та лінійної регресії як найбільш ефективних методів для прогнозування вартості нерухомості. Стаття [5] представляє модель машинного навчання для прогнозування цін на житло. У дослідженні автори використовувалися дані з Шеньчженя, Китай, та комбінували статистичне моделювання та штучні нейронні мережі для прогнозування цін на житло. Автори оцінили продуктивність своєї моделі за допомогою показників середньої абсолютної відсоткової помилки MAPE і середньоквадратичної помилки RMSE, і у висновках показали багатообіцяючі результати. Дослідження [6] спрямоване на прогнозування цін на житло в Мельбурні, Австралія, за допомогою алгоритмів машинного навчання, а саме підтримки векторної регресії та штучних нейронних мереж. Дослідження використовує набір даних за 2008–2016 роки для навчання та тестування моделей і демонструє їхню ефективність у прогнозуванні цін на житло. Стаття має практичне значення для міського планування з використанням технологій машинного навчання. Дослідницька стаття [7] представляє аналіз ефективності різних алгоритмів машинного навчання для прогнозування цін на житло на китайському ринку. Автори обговорюють різні параметри та моделі та оцінюють їх продуктивність за допомогою статистичних оцінок. У статті [8] представлено гібридну регресійну техніку для прогнозування цін на житло, описано методологію та реалізацію техніки гібридної регресії, а також її продуктивність порівняно з іншими моделями регресії.

**Виклад основного матеріалу.** Для задач прогнозування в різних галузях бізнесу можуть використовуватися різноманітні комерційні рішення та платформи. Розглянемо відомі інструменти для предиктивної аналітики великих даних.

*IBM Watson Studio* надає інструменти для розробки моделей машинного навчання та прогнозування, а також інтегровані сервіси та інструменти

для візуалізації великих даних та розгортання прогнозних моделей.

*Azure ML від Microsoft* є хмарним сервісом, який дозволяє створювати, навчати та валідувати моделі машинного навчання, а також надає можливість автоматичного масштабування та розгортання моделей.

*Amazon Forecast* – це хмарний сервіс від Amazon Web Services (AWS), який дозволяє створювати прогнози на основі часових рядів. Використовується для прогнозування попиту, продажів тощо.

*Google Cloud AI Platform* надає засоби для створення, навчання та розгортання моделей машинного навчання на хмарній інфраструктурі Google Cloud. SAS є однією з провідних компаній у сфері аналітики та прогнозування. Програмне забезпечення, таке як *SAS Enterprise Miner*, надає інструменти для створення та валідації моделей прогнозування.

Oracle пропонує рішення для машинного навчання та аналітики, включаючи *Oracle Machine Learning*, що використовується для створення моделей прогнозування на основі даних Oracle.

*RapidMiner* – це платформа для аналізу великих даних та машинного навчання, яка використовується для прогнозування, класифікації та інших аналітичних завдань.

Розглянуті програмні рішення надають інструменти для розробки, навчання та використання моделей прогнозування в різних областях бізнесу. Основним недоліком наявних рішень є висока вартість сервісів для аналітики великих даних, відсутність даних для українського сегменту, що може бути проблемою для практичного використання, зокрема, для прогнозування ринку нерухомості в Україні.

Регресійний аналіз – це статистичний метод, який використовується для дослідження взаємозв'язку між однією залежною та однією чи кількома незалежними (пояснювальними) змінними. Головна мета регресійного аналізу – визначити природу і силу зв'язку між цими змінними. Наприклад, для прогнозування ціни будинку на основі його розміру будується модель лінійної регресії. Модель відповідає прямій лінії до точок даних, щоб зафіксувати основну тенденцію. Розширивши цю лінію, можна оцінити ціну для даного розміру будинку. Лінійна регресія відноситься до категорії регресійних моделей, оскільки вона передбачає числові значення, наприклад, значення ціни на житло. Для навчання моделі лінійної регресії використовується навчальний набір даних, який складається з пар вхідних характеристик (наприклад, квадратних метрів житла) і відповідних кінцевих цільових показників (наприклад, ціна будинку).

Допустимо, що ми маємо:

$m$ – кількість навчальних прикладів;

$x$ – вхідна характеристика, що представляє конкретну властивість будинку, наприклад, його розмір;

$y$ – вихідна ціль, що представляє фактичну ціну будинку;

$i$ – індекс конкретного навчального прикладу.

Конкретний приклад навчання позначається як показано у формулі (1).

$$(x^i, y^i) \tag{1}$$

Отже навчальний набір можна записати, як у формулі (2).

$$\{(x^1, y^1),(x^2, y^2),\ldots,(x^m, y^m)\} \tag{2}$$

У рівнянні (2) $(x^i, y^i)$ представляє певний будинок у наборі даних, а $m$ позначає кількість прикладів для навчання.

Під час навчання моделі алгоритм навчання створює функцію $f$, яка приймає вхідну ознаку $x$ і виводить оцінку або прогноз $y$. У простій лінійній регресії функція $f(x)$ визначається за формулою (3).

$$f_{w,b}\left(x^{(i)}\right) = wx^{(i)} + b \tag{3}$$

В цій формулі $w$ і $b$ – це числові значення, які визначають нахил і перетин ліній відповідно. Важливим аспектом лінійної регресії є функція втрат, яка вимірює ефективність моделі шляхом кількісного визначення різниці між прогнозованими значеннями та фактичними цільовими показниками. Ця концепція широко застосовується в машинному навчанні та відіграє життєво важливу роль у навчанні моделей штучного інтелекту.

Лінійна регресія є цінним інструментом для прогнозування цін на житло на основі конкретних характеристик. Ринок нерухомості відомий своєю нестабільністю та непередбачуваністю. Точні прогнози вартості нерухомості мають вирішальне значення для інвесторів, забудовників і власників житла. Традиційні методи прогнозування, такі як лінійна регресія, мають обмеження у відображенні складних взаємозв'язків між різними факторами, які впливають на ціну нерухомості. Дослідження пропонує більш складну модель на основі регресійного аналізу для усунення цих обмежень. Спочатку у дослідженні визначатимуться ключові фактори, які суттєво впливають на вартість нерухомості. Ці фактори включають розташування, розмір нерухомості, зручності, економічні показники та демографічні тенденції. Аналізуючи історичні дані та проводячи статистичні тести,

будуть визначені найбільш впливові змінні для включення в регресійну модель. Після визначення факторів впливу розроблятиметься регресійна модель. Ця модель використовуватиме історичні дані для встановлення зв'язків між незалежними змінними (факторами впливу) і залежною змінною (вартістю нерухомості). Потім модель буде перевірено з використанням додаткових (тестувальних) даних для забезпечення її точності та надійності.

Програмна реалізація запропонованого методу прогнозування

Для даного дослідження було обрано мову програмування Python, найпопулярнішу мову програмування для аналізу даних та машинного навчання. Для роботи з даними використовувались бібліотеки Pandas та NumPy. Бібліотека Pandas надає структури даних високого рівня, такі як DataFrame, що дозволяють ефективно маніпулювати даними та аналізувати їх. Pandas використовується для читання та запису даних, фільтрації, групування та обробки даних перед подальшим аналізом. Бібліотека NumPy надає масиви та функції для ефективної роботи з числовими даними, особливо є корисною для векторизованих операцій, виконання числових операцій та роботи з масивами даних. Для візуалізації даних використовуються бібліотеки Matplotlib та Seaborn. Для моделювання та машинного навчання використано високорівневу бібліотеку машинного навчання Scikit-learn, яка містить реалізації багатьох алгоритмів машинного навчання. Для розроблення програмного методу прогнозування вартості нерухомості у даному дослідженні використано хмарне середовище Google Colab (рис. 1).

Google Colab – це хмарне середовище для виконання коду на мові програмування Python, яке надає доступ до віртуальних машин Google з можливістю використання графічних процесорів GPU для прискорення обчислень, що особливо корисно для задач машинного навчання. Для середовища Google Colab доступні віртуальні машини з різними характеристиками пам'яті та обчислювальною потужністю. Зазвичай, базова конфігурація може включати в себе 13–14 ГБ оперативної пам'яті і процесор з 2–4 ядрами. У деяких випадках, Colab може також надавати доступ до графічних процесорів (GPU) від Nvidia, таких як Tesla K80, Tesla T4 або P100, які забезпечують додаткові обчислювальні ресурси та можуть прискорити виконання завдань, пов'язаних із штучним інтелектом або машинним навчанням. Вибір ознак прогнозної моделі виконується за допомогою кореляційного аналізу, визначення важливості ознак за допомогою моделей та рекурсивного вилучення ознак RFE (Recursive Feature Extraction). Кореляційний аналіз оцінює взаємозв'язок між ознаками та вибирає ті, які мають найбільший вплив на цільову змінну. RFE використовується для ранжування ознак та вилучення їх послідовно відповідно до їх важливості для моделі, ефективно визначає оптимальну кількість ознак для покращення точності прогнозної моделі. Інженерія ознак виконується за допомогою створення взаємодій між ознаками, створення ознак на основі доменного знання, агрегації статистики і кодування категоріальних ознак. Множинна лінійна регресія використовується для врахування різних факторів, таких як площа, кількість кімнат, розташування тощо. Регресія на основі ансамблевих методів (наприклад, випадковий ліс або градієнтний бустинг) використовує комбінації декількох моделей для покращення точності та стабільності прогнозів. Така стратегія може допомогти уникнути перенавчання та покращити прогнози. При використанні регресійних моделей важливо ретельно вибирати та налаштовувати ознаки, враховувати нелінійні взаємозв'язки та коригувати модель відповідно до особливостей даних. Також слід уникати пере-

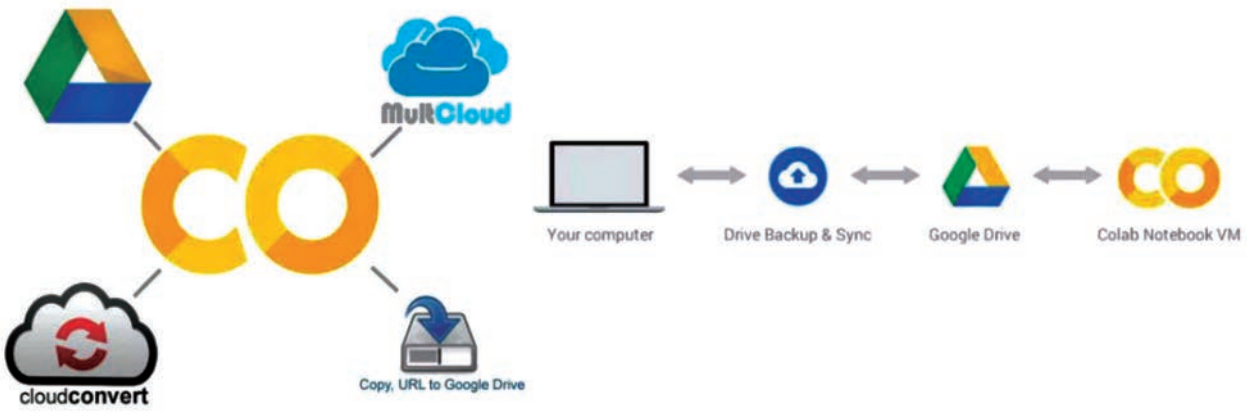


**Рис. 1. Використання хмарного середовища Google Colab**

навчання, використовуючи методи регуляризації (наприклад, L1 або L2 регуляризацію). Регресійні моделі можуть бути ефективним інструментом для прогнозування вартості нерухомості, особливо при правильному виборі та обробці ознак.

Оптимізація алгоритмів та коду може значно покращити швидкодію прогностичних моделей. Використання векторизованих операцій замість циклів може допомогти прискорити обчислення в багатьох випадках. Також варто уникати циклів у Python, особливо при роботі з великими об'ємами даних. Для розпаралелювання обчислень та використання багатоядерних систем можна використовувати можливості паралельного обчислення, такі як бібліотека Multiprocessing у Python. Використання бінарних форматів для збереження та завантаження даних, такі як HDF5 або Parquet, може бути ефективнішим за використання текстових форматів. Варто уникати зайвого використання пам'яті, особливо при роботі з великими наборами даних. Доцільно підібрати гіперпараметри для отримання більшої ефективності та швидкодії, а також використовувати такі методи оптимізації гіперпараметрів, як Grid Search або Random Search. Важливим аспектом ефективності оптимізації є час виконання коду та результати прогнозування моделі. Для перевірки оцінки точності прогнозування моделі у даному дослідженні використано середньо-квадратичну помилку RMSE та коефіцієнт детермінації R-squared. Достовірність моделі оцінюється за допомогою валідації моделі та перевірки на мультиколінеарність. Дані розділяються на тренувальний та тестовий набори для перевірки стабільності моделі. Оцінюється наявність мультиколінеарності серед факторів, що впливають на ціни на нерухомість, щоб уникнути спотворень у моделі. Стабільність прогнозів оцінюється за допомогою аналізу змін в часі. Порівняння з іншими прогнозними моделями відбувається за рахунок порівняння результатів точності після їх тестування.

У даному дослідженні використовується датасет «Linear Regression – House Price Predictions», обсягом 514 КВ, який містить дані про продажі нерухомості в США за три місяці в період травень-липень 2014 року. Досліджуваний датасет містить 4601 запис та колонки: date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition. Колонка date відповідає за дату, коли було отримано інформацію про нерухомість, price – ціна нерухомості, bedrooms – кількість спалень, bathrooms – кількість ваних кімнат, sqft_living – житлова площа, sqft_lot – площа земельної ділянки, floors – кількість поверхів, waterfront вказує, чи має нерухомість вихід до води, наприклад озера, ріки, океану (1 – є вихід до води, 0 – вихід до води відсутній), view вказує на панорамний вид з квартири чи будинку та є категорійним атрибутом, condition вказує на загальний стан будівлі (1 – екстримально поганий стан, 2 – дуже поганий стан, 3 – середній стан, 4 – добрий стан, 5 – відмінний стан). Спочатку ми досліджуємо та виконуємо попередню обробку набору даних, імпортуємо необхідні бібліотеки для роботи з даними, візуалізації та машинного навчання.

```
1  import pandas as pd
2  import matplotlib.pyplot as plt
3  import seaborn as sns
4  from sklearn.model_selection import train_test_split
5  from sklearn.linear_model import LinearRegression
6  from sklearn.metrics import mean_squared_error, r2_score
```

Дані завантажуються з CSV-файлу в об'єкт DataFrame за допомогою бібліотеки Pandas та записуються у змінну *df*, потім проводиться огляд даних, розрахунок основних статистичних характеристик та перевірка відсутності пропущених значень.

```
8  # Load the dataset from CSV
9  df = pd.read_csv('house_data.csv')
10
11 # Exploratory Data Analysis (EDA)
12 # Let's take a quick look at the first few rows of the dataset
13 print(df.head())
14
15 # Summary statistics of the dataset
16 print(df.describe())
17
18 # Check for missing values
19 print(df.isnull().sum())
```

Потім створюємо кореляційну теплову матрицю для аналізу взаємозв'язків між характеристиками у вигляді кольорової картинки (рис. 2).

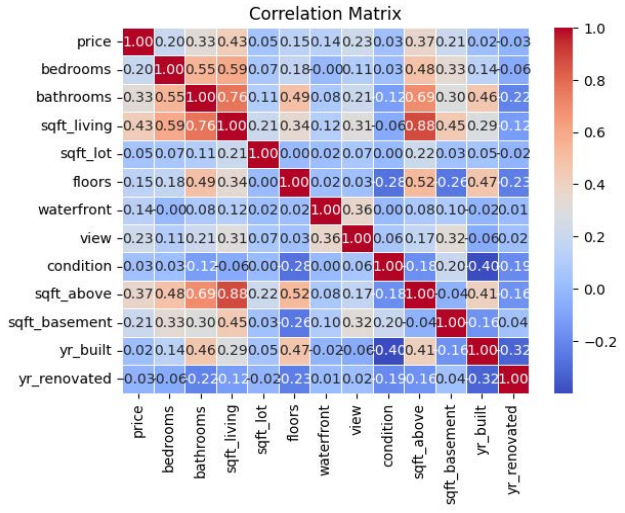


**Рис. 2. Матриця кореляції для аналізу взаємозв'язків між різними ознаками**

У рядках 28–29 обираємо ознаки та цільову змінну для подальшого їх використання у прогнозній моделі.

```
21 # Correlation matrix to understand feature relationships
22 correlation_matrix = df.corr()
23 sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
24 plt.title("Correlation Matrix")
25 plt.show()
```

У рядку 32 дані розділяються на навчальний і тестовий набори для ефективного тренування та перевірки моделі.

```
31  # Splitting the dataset into training and testing sets
32  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Будуємо модель лінійної регресії, використовуючи клас *LinearRegression* з бібліотеки *scikit-learn* та метод *.fit()* для навчання моделі на навчальних даних.

```
37  # Building the Linear Regression Model
38  model = LinearRegression()
39
40  # Fitting the model on the training data
41  model.fit(X_train, y_train)
```

*X_train* представляє навчальні ознаки, а *y_train* – цільову змінну (вартість нерухомості в нашому випадку). Під час навчання модель знаходить оптимальні параметри (ваги) для лінійної регресії, які найкраще відображають зв'язок між ознаками і цільовою змінною.

Побудуємо коефіцієнти регресії. Для цього використовуємо *X.columns* – атрибут DataFrame, який повертає імена стовпців в датафреймі *X* (в даному випадку, це назви факторів, які були вибрані для використання в моделі) та *model.coef_* – атрибут, який містить коефіцієнти регресії, які модель призначила кожному фактору (кожен коефіцієнт показує, на скільки зміниться цільова змінна при зміні на одиницю відповідного фактора).

```
42  # Building regression coefficients
43  coefficients = pd.DataFrame({'Feature': X.columns, 'Coefficient': model.coef_})
44  print(coefficients)
```

Таким чином, створюємо новий об'єкт DataFrame, де об'єднуємо назви факторів (*X.columns*) і відповідні їм коефіцієнти регресії (*model.coef_*) у два стовпці: 'Feature' та 'Coefficient'.

| | Feature | Coefficient |
|---|---|---|
| 0 | bedrooms | -56797.399509 |
| 1 | bathrooms | -8727.410252 |
| 2 | sqft_living | 285.272386 |
| 3 | sqft_lot | -0.564912 |
| 4 | floors | 38931.386966 |
| 5 | waterfront | 407431.460385 |
| 6 | view | 55854.014776 |
| 7 | condition | 56771.209365 |

Проводимо оцінку моделі, застосовуючи навчену модель (*model*), що є об'єктом лінійної регресії, до тестового набору ознак (*X_test*). Модель використовує навчені ваги для здійснення прогнозів вартості нерухомості на основі тестових ознак. Результат

```
47  # Model Evaluation
48  y_pred = model.predict(X_test)
49
50  # Mean Squared Error and R-squared for model evaluation
51  mse = mean_squared_error(y_test, y_pred)
52  r2 = r2_score(y_test, y_pred)
53
54  print("Mean Squared Error:", mse)
55  print("R-squared:", r2)
```

У рядку 51 використовуємо функцію *mean_squared_error* з бібліотеки Scikit-learn для обчислення середньоквадратичної помилки MSE між реальними значеннями цільової змінної (*y_test*) і прогнозованими значеннями (*y_pred*). У рядку 52 використовуємо функцію *r2_score* для обчислення коефіцієнта детермінації R-squared між реальними значеннями (*y_test*) і прогнозованими значеннями (*y_pred*). Результати оцінки MSE і R-squared зберігаються у змінних *mse* і *r2*.

Отримуємо середню квадратичну помилку 94953,98 та R-квадрат 0,86233518995632512. Далі отримаємо прогнози та візуалізуємо їх.

Рядок 63 коду дозволяє створити точковий графік, де по горизонтальній осі розташовані фактичні ціни (*y_test*), а по вертикальній осі – прогнозовані ціни (*y_pred*).

```
63  plt.scatter(y_test, y_pred)
64  plt.xlabel("Actual Prices")
65  plt.ylabel("Predicted Prices")
66  plt.title("Actual Prices vs. Predicted Prices")
67  plt.show()
```

У рядках 64–65 ми описуємо позначення осей графіку. У рядку 66 вказуємо заголовок графіку. Рядок 67 відповідає за відображення графіку. Цей графік дозволяє порівняти прогнозовані та фактичні ціни на нерухомість.

```
69  # Creating a residual plot to check the model's performance
70  residuals = y_test - y_pred
71  plt.scatter(y_test, residuals)
72  plt.axhline(y=0, color='red', linestyle='--')
73  plt.xlabel("Actual Prices")
74  plt.ylabel("Residuals")
75  plt.title("Residual Plot")
76  plt.show()
```

У рядку 70 розраховуємо залишкові (*residual*) значення, що представляють різницю між фактичними цінами та прогнозованими цінами.

У рядку 71 створюємо точковий графік, де по горизонтальній осі розташовані фактичні ціни, а по вертикальній осі – залишкові значення.

У рядку 72 додаємо горизонтальну лінію на рівні 0 червоного кольору для визначення нульового рівня залишкових значень. У рядках 73–75 позначаємо осі графіку, вказуємо заголовок графіку.

Рядок 76 відповідає за відображення графіку залишкових значень, який допомагає оцінити, наскільки добре модель враховує варіацію в даних та чи є які-небудь систематичні відхилення.

```
78  # Using the trained model to make predictions on new data and visualize the results
79  new_data = [[3, 2, 1500, 4000, 1, 0, 0, 3]]
80  predicted_price = model.predict(new_data)
81
82  print("Predicted Price:", predicted_price[0])
```

У рядку 79 створюємо новий набір даних для прогнозування ціни. Вказуємо значення для кількості спалень, ванних кімнат, площі, розміру ділянки, кількості поверхів, наявності водойми, виду та стану.

У рядку 80 застосовуємо навчену модель для прогнозу ціни на нових даних. Рядок 82 виводить прогнозовану ціну. Отримавши візуалізацію (рис. 3), бачимо що на залишковій ділянці модель видає ближчі до реальності результати, аніж на тестовому наборі даних. Графік Actual Prices vs. Predicted Prices дозволяє порівняти прогнозовані та фактичні ціни на нерухомість. Графік Residual Plot допомагає оцінити, наскільки добре модель враховує варіацію в даних та чи є які-небудь систематичні відхилення.

Застосування розробленої моделі в практичних умовах має ряд перспектив на ринку нерухомості України. Інвестори можуть використовувати прогнозні моделі для оптимізації своїх інвестицій у нерухомість, визначаючи об'єкти з найбільшим потенціалом зростання вартості. Забудовники можуть використовувати прогнози для оптимізації свого ресурсного планування, визначаючи ринкові тенденції та попит на певні типи нерухомості.

Областю застосування запропонованого програмного методу є прогнозування цін на житло

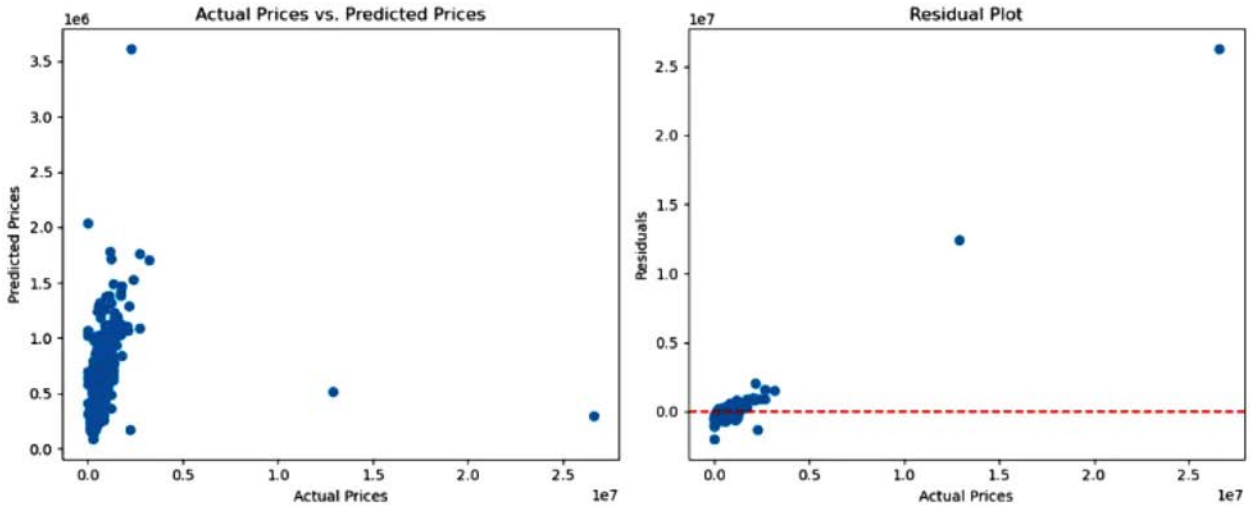


**Рис. 3. Прогнози моделі на тестовому та на залишковому наборах даних**

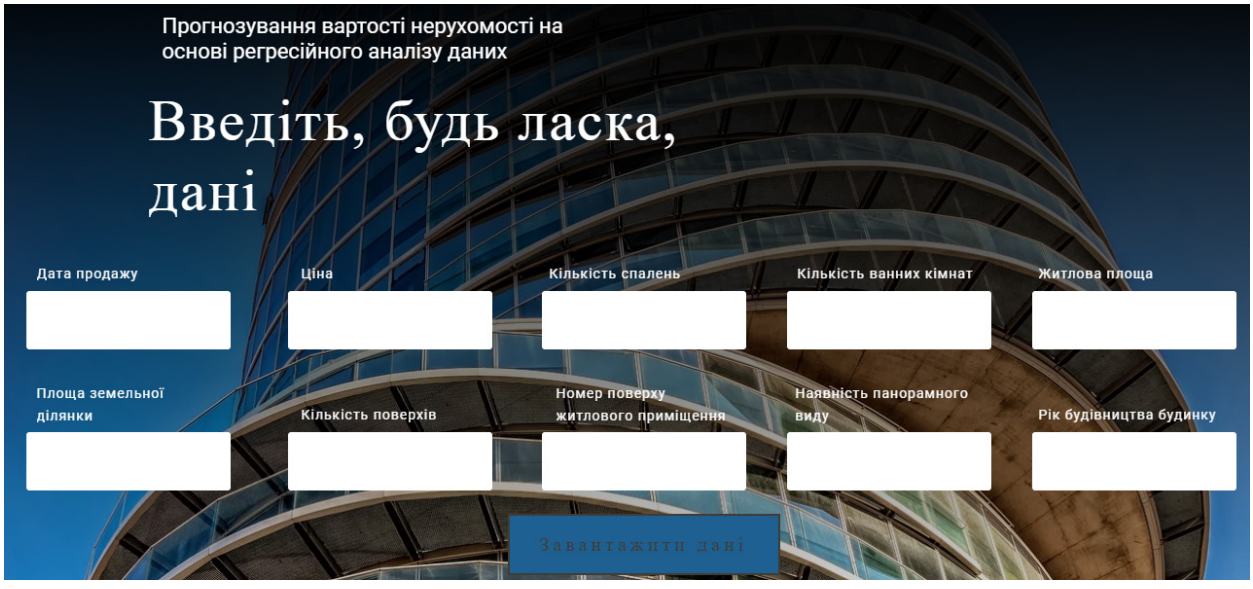


**Рис. 4. Графічний інтерфейс користувача для збору та аналізу даних про продажі нерухомості в Україні**

в різних регіонах України, комерційну нерухомість чи орендні ставки; прогнозування ринкових тенденцій для прийняття інвестиційних рішень; визначення вартості страхового покриття для нерухомості; аналіз впливу географічного розташування нерухомості на ціни; визначення оптимального часу для купівлі чи продажу нерухомості. Різні країни та регіони можуть мати різні ставки податку на нерухомість, які також впливають на вартість власності та інвестиційну привабливість.

Авторами було запропоновано і створено графічний інтерфейс користувача для збору та аналізу даних про продажі нерухомості України, який містить поля для введення та збереження статистичних даних про продажі житла різного типу (рис. 4).

У даному дослідженні було проаналізовано використання різних методів регресійного аналізу даних для прогнозування вартості нерухомості для обраного датасету, оскільки прогнозування вартості нерухомості є ключовим етапом у сфері нерухомості та фінансового планування. Лінійна регресія може використовуватися для моделювання простих взаємозв'язків між однією або декількома незалежними змінними та залежною змінною, такою як ціна нерухомості. Метод градієнтного бустингу, наприклад, застосовується для покращення точності прогнозування, враховуючи взаємозв'язки та нелінійність в даних. Також, регресійні моделі, побудовані за допомогою нейронних мереж, можуть ефективно моделювати складніші залежності вартості нерухомості враховуючи багатовимірні аспекти.

Регресія Lasso (Least Absolute Shrinkage and Selection Operator) є типом лінійної регресії, де сума значень коефіцієнтів штрафується, щоб уникнути помилок передбачення. Виходячи з досліджень спільноти Kaggle «Machine Learning Approach for House Price Prediction» бачимо, що така модель регресії дає точніші результати на 1%. Дослідження показали, що алгоритм машинного навчання SVM (Support Vector Machines) виконує свою функцію передбачення значно гірше за класичну лінійну регресію, оскільки різниця становить 58% точності. Алгоритм випадкового лісу (Random Forest) використовує кілька дерев рішень для прогнозування результату. Згідно з проведеного дослідження, випадковий ліс є найкращим типом регресії для передбачень, оскільки точність передбачень є на 11% вищою порівняно з використанням звичайної лінійної регресії. XGBoost – це алгоритм машинного навчання, який може обробляти складні взаємодії функцій та ефективно фіксувати нелінійні зв'язки, що забезпечує точніші прогнози. Методи регуляризації, які використовуються в XGBoost, допомагають запобігти перенавчанню та підвищити ефективність моделі. Точність цього алгоритму вища на 9% від точності класичної лінійної регресії. Проте налаштування гіперпараметрів XGBoost може бути складним завданням, і неправильний вибір параметрів може призвести до перенавчання чи недонавчання моделі, що може вплинути на її ефективність.

**Висновки та подальша робота.** У даному дослідженні запропоновано програмний метод прогнозування вартості нерухомості на основі регресійного аналізу даних та методів машинного навчання з перспективою впровадження даного методу для збору та аналізу даних на ринку нерухомості України.

Подальші дослідження авторів передбачають створення відкритої інформаційної системи для збору даних про нерухомість в Україні, врахування в регресійній моделі специфічних економічних факторів, які можуть впливати на ціни нерухомості в регіонах України, таких як попит і пропозиція, процентні ставки, інфляція та економічне зростання. Фінансові аспекти, включаючи варіанти фінансування та іпотечні ставки, а також державна політика відіграють вирішальну роль у прогнозуванні вартості нерухомості в Україні, що також потрібно враховувати і аналізувати при побудові прогнозних моделей на основі регресійного аналізу даних.

**Список літератури:**

1. Plakandaras V., Gupta R., Gogas P., Papadimitriou T., 2015. Forecasting the U.S. real house price index, *Economic Modelling, Elsevier*, vol. 45(C), pages 259–267. DOI: 10.1016/j.econmod.2014.10.050.

2. Li Y., Leatham D. Forecasting housing prices: dynamic factor model versus LBVAR model. *2011 Annual Meeting, Pittsburgh, Pennsylvania, Agricultural and Applied Economics Association*, July 24-26, 2011, 103667. DOI: 10.22004/ag.econ.103667.

3. Beracha E., Wintoki B. Forecasting residential real estate price changes from online search activity. *Journal of Real Estate Research,* 35(3). July 2013. DOI: 10.1080/10835547.2013.12091364.

4. M. Jagan Chowhaan, D. Nitish, G. Akash, Nelli Sreevidya and Subhani Shaik. Machine Learning Approach for House Price Prediction. *Asian Journal of Research in Computer Science,* Volume 16, Issue 2, pages 54–61, 2023. DOI: 10.9734/ajrcos/2023/v16i2339.

5. Fan C., Cui Z., Zhong X. House Prices Prediction with Machine Learning Algorithms. *Proceedings of the 2018 10th International Conferenceon Machine Learning and Computing – ICMLC 2018*. doi:10.1145/3195106.3195133.

6. Phan T.D. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (ICMLDE) 2018*. doi:10.1109/icmlde.2018.00017.

7. Mu J., Wu F., Zhang A. Housing Value Forecasting Based on Machine Learning Methods. *Abstract and Applied Analysis 2014*; 2014:1–7. doi:10.1155/2014/648047.

8. Lu S., Li Z., Qin Z., Yang X. A hybrid regression technique for house prices prediction. *2017 IEEE International Conference onIndustrial Engineering and Engineering Management (IEEM) 2017*. doi:10.1109/ieem.2017.8289904.

**Oleshchenko L.M., Trushyna D.V. PREDICTIVE SOFTWARE METHOD FOR REAL ESTATE VALUATION USING MACHINE LEARNING AND REGRESSION ANALYSIS**

*The article considers the features of the software implementation of the regression model for forecasting the value of real estate using machine learning methods. To select a housing price forecasting model, an analysis of publications and existing software solutions for real estate price forecasting was conducted. In this research, the Python programming language, Pandas, Matplotlib, Seaborn, Scikit-learn, and NumPy libraries were used for the practical implementation of the proposed software method. The "Linear Regression – House Price Predictions" dataset, with a volume of 514 KB, containing 4601 records, was chosen for the study of building a regression model. The model was trained, its accuracy was evaluated and compared with other forecasting methods using machine learning technologies.*

*The article provides a study of the areas of application of the model for forecasting the value of real estate based on regression analysis of data, the formation of software requirements, the choice of a programming language, the development of a structure and the training of an algorithm. The technological approaches covered include data processing and preparation, feature selection, regression modeling, model training and evaluation, integration of artificial intelligence and machine learning, and model validation and optimization.*

*As a result of the research, a software method was created for forecasting the value of real estate based on regression analysis of data, which provides a graphical display of a list of factors with a quantitative assessment of their impact on the value of real estate. The advantages of this study are important from the point of view of the software implementation of the method of regression analysis of data and the possibility of future use of the proposed method for forecasting the real estate market in Ukraine. By developing a model based on regression analysis, real estate professionals will have a more reliable tool for predicting real estate values. This will allow them to make informed decisions about investments, pricing strategies and development projects. In addition, property owners will have a better understanding of the factors that affect their property's value, allowing them to make informed decisions about home repairs and improvements.*

***Key words:*** *programming technologies of artificial intelligence systems, Python programming language, forecasting, machine learning, real estate market, big data, data processing, data regression analysis.*