УДК 004.415

ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ЦІН НА ЖИТЛО

Труба Олександр

Науковий керівник: доктор технічних наук, професор Казачков І. В.

Ніжинський державний університет імені Миколи Гоголя, м. Ніжин, Україна

У статті проілюстровано застосування методів машинного навчання з метою аналізу та прогнозу цін на ринку нерухомості. Основна увага приділена ретельному аналізу вхідних даних та розгляду різних варіантів регресійних моделей. У результаті проведених досліджень було розроблено статистичну модель для прогнозування цін на житлову нерухомість, використовуючи метод випадкового лісу, як найкращу модель. Процес створення моделі включав етапи імпорту бібліотек та модулів, завантаження даних з датасету, аналізу даних, їх очищення та середньої статистичної оцінки. Декілька методів машинного навчання було ретельно вивчено з використанням середньоквадратичної похибки для визначення оптимального підходу. Масив даних, отриманий із відкритого ресурсу – сайту продажу та оренди нерухомості "Лун", ефективно використаний для аналізу, зокрема завдяки кластеризації за параметрами, такими як площа, розміщення, рік побудови тощо. Підготовка та тестування фінальних моделей проводилися на реальних даних, що продемонструвало відмінні результати в прогнозуванні цін. Отриманий результат моделювання був ефективно перевірений на реальних об'єктах нерухомості, підтверджуючи точність роботи моделі на рівні 87%. Високі показники свідчать про правильність конструкції моделі та доцільність використання програмних рішень для її втілення. Зазначена модель легко може застосовуватися для аналізу подібних масивів даних.

Ключові слова: машинне навчання, регресійні моделі, ціна на нерухомість, випадковий ліс, візуалізація даних, прогнозування.

Application of machine learning for analysis and prediction of real estate prices

O. Truba

Scientific supervisor: Doctor of Technology, Professor Kazachkov I. V.

Nizhyn Gogol State University, Nizhyn, Ukraine

The article illustrates the application of machine learning methods to analyze and forecast prices in the real estate market. The main attention is paid to a thorough analysis of the input data and consideration of various variants of regression models. As a result of the research, a statistical model

for forecasting residential real estate prices was developed using the random forest method as the best model. The process of creating the model included the steps of importing libraries and modules, loading data from the dataset, analyzing the data, cleaning it, and performing a statistical mean. Several machine learning methods were thoroughly evaluated using root mean square error to determine the best approach. The dataset obtained from an open resource, the real estate sales and rental website Lun, was effectively used for analysis, in particular through clustering by parameters such as area, location, year of construction, etc. The final models were trained and tested on real data, which demonstrated excellent results in price forecasting. The modeling result was effectively validated on real estate objects, confirming the accuracy of the model at 87%. These high results indicate the correctness of the model design and the feasibility of using software solutions for its implementation. The model can be easily applied to analyze similar data sets.

Key words: machine learning, regression models, real estate price, random forest, data visualization, forecasting..

Постановка проблеми. Методи машинного навчання та основи штучного інтелекту широко використовуються в різних галузях, включаючи науку, технології та бізнес, де важливо аналізувати великі обсяги даних за допомогою постійних алгоритмів взаємодії. Засадовою частиною машинного навчання є правильна інтерпретація даних та їх використання для вирішення конкретних завдань.

Однією з актуальних проблем сучасності є аналіз формування цін на житло, особливо в умовах кризових ситуацій. Динаміка цін на ринку нерухомості в таких умовах суттєво відрізняється від стабільних економічних періодів. Ціноутворення на ринку нерухомості базується на взаємодії мікроекономічних факторів, таких як попит і пропозиція, які залежать від макроекономічних показників розвитку країни, регіону та міста. Аналіз цих параметрів можна виконувати за допомогою різних моделей, зокрема, регресійних, які застосовуються для передбачення цільових змінних у безперервній шкалі. Регресійні моделі є ефективним інструментом для розв'язання різних завдань у науці та інформаційній галузі, таких як вивчення взаємозв'язків між змінними, оцінка тенденцій та прогнозування. Прикладом такого застосування є прогнозування за допомогою моделей, що використовують активні оголошення, які відображають поточний стан ринку, як вхідний набір даних.

Аналіз останніх досліджень і публікацій. Аналіз ринку нерухомості, як самостійна робота так і для підготовки моделей машинного навчання, є актуальним предметом досліджень. Тема ціноутворення на ринку нерухомості розглядаються в роботах А. М. Асаула [1] та Ю. М. Манцевича [2]; застосування нейронних мереж для прогнозування цін на нерухомість розглядаються в роботах В. С. Григорківа [3], Н. О. Філіпчука [3], О. М. Ярошенко [3], В.А. Вороніна [4]. Варто зазначити, що вибрані методи моделювання повністю залежать від мети дослідження та структури даних, а також через різноманітність завдань досліджень. Тому на даний момент відсутній єдиний концептуальний підхід до аналізу та прогнозування цін на нерухомість в межах міста.

Мета статті. Розробка концептуальної моделі для прогнозування цін на об'єкти нерухомості в межах одного міста за допомогою методів машинного навчання, які враховують ключові чинники, що впливають на формування цін.

Виклад основного матеріалу дослідження.

Питання придбання житла є завжди актуальним і водночас завжди визначається складністю, вимагаючи врахування різноманітних факторів та нюансів. Вибір нерухомості зазвичай обумовлюється увагою до її стану, планування та дизайну, проте в центрі уваги завжди перебуває ціновий аспект. У зусиллях оцінки нерухомості ми систематично намагаємося узагальнити всі аспекти та визначити, наскільки ціна відповідає реальному становищу.

Для побудови ефективної моделі прогнозування необхідно мати обширний обсяг даних. І тому для побудови моделей булоа обрано місто Київ як об'єкт дослідження, вважаючи, що тут розташовано найбільше оголошень про продаж житла, зокрема квартир. В рамках дослідження розглядаються різні методи побудови моделей, і на завершальному етапі вибирається той, який виявляє найменшу похибку відповідно до обраної метрики, такої як RMSE (середньоквадратична похибка).

Початок процесу передбачає підготовку даних, для якої використовується сайт "Лун", як джерело інформації, оскільки він одним з найпопулярніших ресурсів для продажу житла. Такий підхід дозволяє отримувати актуальні дані та коригувати модель відповідно до їхньої актуальності. Почнемо з етапу завантаження даних, а далі використаємо існуючі бібліотеки для аналізу.

Мені вдалося отримати приблизно 25 тисяч записів. Приклад даних представлено на рис.1. Варто зазнчити, що я зробив додаткову фільтрацію, щоб відкинути некоректні дані. Наприклад, всі квартири ціною більше за 1 мільйон доларів або будинку, які побудовані раніше, ніж 1920 рік. Звісно, існує ймовірність, що так можна втрати певну частину інформації, але ця частка досить мала і не зашкодить подальшому аналізу.

	area_kitchen	area_living	area_total	built_year	ceiling_height	floor	floor_count	district	heating_system_name	house_type_name
0	7.0	32.9	47.8	1971.00000	2.65	7.0	9.0	Святошинський	centralized	NaN
1	13.8	21.3	59.4	2002.00000	2.70	8.0	18.0	Святошинський	autonomous	special_project
2	11.9	50.6	88.8	2003.00000	2.75	3.0	11.0	Святошинський	centralized	special_project
3	14.0	16.0	39.6	2002.81352	NaN	5.0	6.0	Святошинський	centralized	NaN
4	8.0	45.0	74.0	1992.00000	2.70	2.0	16.0	Святошинський	centralized	series_kt
5	13.3	53.2	93.2	2005.00000	2.70	14.0	22.0	Святошинський	centralized	special_project
6	9.0	16.0	37.0	2002.81352	NaN	19.0	25.0	Святошинський	centralized	NaN
7	13.0	53.0	87.0	2018.00000	2.70	11.0	17.0	Святошинський	centralized	special_project
8	12.0	52.0	91.0	2003.00000	2.75	10.0	11.0	Святошинський	centralized	special_project
9	16.0	40.0	83.0	2012.00000	2.70	12.0	17.0	Святошинський	centralized	special_project

Рис. 1 Представлення вхідних даних

Для кожного запису враховується 16 параметрів. Це такі як: площа кухні, площа житлових приміщень, загальна площа квартири, дата побудови будинку, висота стелі, поверх, загальні кількість поверхів в будинку, район, тип опалення, тип будинку, довгота, широта, ціна, кількість кімнат, матеріал стін та чи відноситься квартира до житлового комплексу.

Дослідимо наші дані трохи детальніше. Для цього побудуємо гістограми для основних характеристик. Як можемо побачити на рис. 2, що понад 8000 будинків були побудовані в 2000-х роках. Деякі характеристики виглядають прямолінійними, наприклад поверх квартири, побачимо найбільше будинків десь до 10 поверху, а потім іде вже спад. Якщо є проблема з обмеженням, ми можемо або зібрати

відповідні значення для обмежень, або вилучити ці райони з наборів даних. Ми також бачимо, що не всі ознаки мають однаковий масштаб, і багато ознак мають "хвіст", тобто простягаються на багато правіше від медіани, ніж лівіше. Це можна побачити по такому параметру як загальна площа, що може свідчити про дані з потенційною похибкою – неправильно вказана площа.

```
# Lets look at the distribution of all the numeric variables
%matplotlib inline
import matplotlib.pyplot as plt
housing.hist(bins=50, figsize=(20,15),)
plt.show()
```

✓ 5.4s Pytho

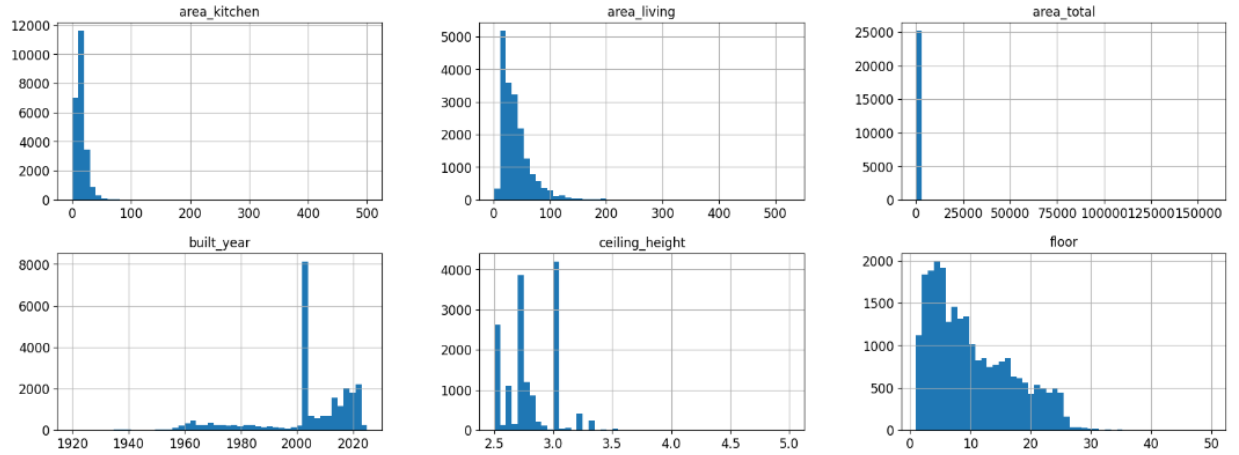

Рис. 2 Представлення даних у вигляді гістограм

Ще одним цікавим представленням є щільності забудови, який представлений на рис. 3. Варто також враховувати, що це може трохи заплутати, оскільки багатоповерхові будинки матимуть більше квартир, ніж будинки з невеликою кількістю поверхів. Тому історичні райони, де будинки є досить низькими будуть відображені, як малозаселенні. Але в загальному, це гарний показник, щоб зрозуміти де проживає велика кількість людей. Тому цей графік можна охарактеризувати, як щільність заселення.

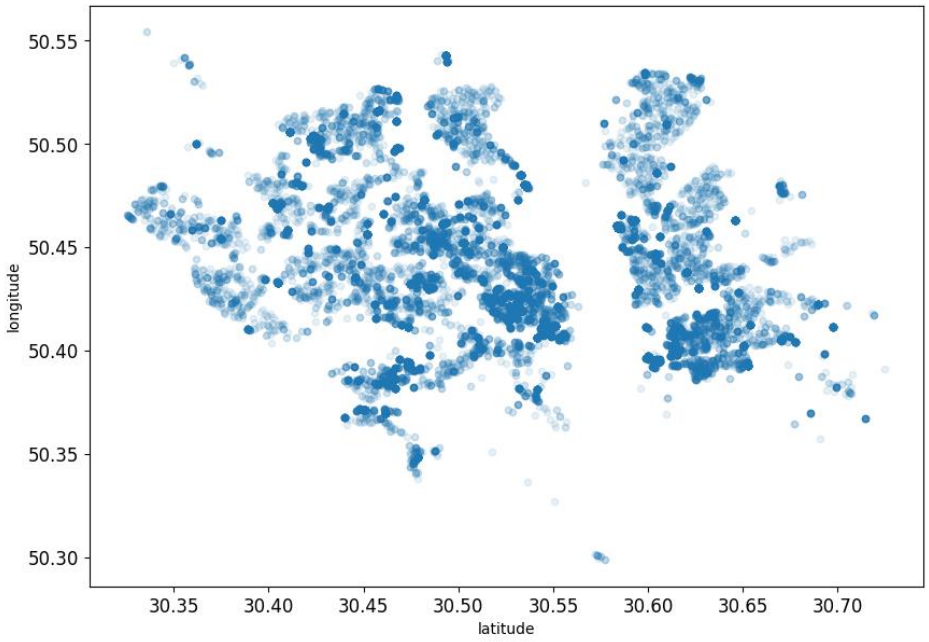

Рис. 3 Відображення щільності забудови

Також побудуємо кореляційну матрицю для характеристик житла. Як можемо бачити на рис. 4, житлова площа, загальна кількість кімнат та висота стелі – це 3 найкращі змінні з точки зору кореляції з нашою цільовою змінною. Для мене це трохи диво, бо мені здавалося, що це не мало би критично впливати на ціну. Можливим поясненням цього є той факт, що більшість квартир з високою стелю будуться в преміальних та вище будинках, які мають зовсім інший сегмент цін, ніж будинку комфорт класу.

```
corr_matrix = housing.corr(numeric_only=True)
corr_matrix["price"].sort_values(ascending=False)
```

✓ 0.1s

```
price                 1.000000
area_living           0.685980
room_count            0.544705
ceiling_height        0.516587
area_kitchen          0.509232
residential_complex   0.212004
built_year            0.206675
floor_count           0.124506
floor                 0.109823
latitude              0.049020
area_total            0.033530
longitude            -0.158131
Name: price, dtype: float64
```

Рис. 4 Кореляційна матриця

Тепер перейдемо до підготовка даних для машинного навчання. Це одна з найважливіших частин, оскільки від того, наскільки добре ми виконаємо цей крок, залежатиме наш результат. Перш ніж виконувати будь-яку функціональну інженерію, ми розділимо набори даних на навчальні та тестові, причому 80% даних підуть на побудову моделі, а 20% – на тестування моделі.

Важливим моментом є перетворення характеристик, які мають значення категорій, у зручний для навчання, формат. Для цього ми використаємо одну техніку гарячого кодування. Для кожного значення категорії створюється одна бінарна характеристика. Для приладу візьме характеристику район. У даному випадку нова характеристика дорівнює 1, якщо значення категорії – район для даної квартири, і 0 для всіх інших районів. У такому випадку лиш одна характеристика буде дорівнювати 1 (гарячий), тоді як інші будуть дорівнювати 0 (холодний). У результаті матимемо одну додаткову характеристику на кожне можливе значення з основної категорії [6].

Провівши всі маніпуляції з вхідним даними, отримаємо навчальний набір даних, який містить 20164 рядків і 47 змінних. Тепер можна перейти до етапу навчання моделі.

Для визначення кращої моделі було обрано декілька методів машинного навчання: лінійна регресія, дерева рішень та випадковий ліс. Після проведення досліджень було отримано результати, наведені у таблиці 1. Метод випадковий ліс показав найменше відхилення на тестових даних, тому використаємо цю модель для перевірки на тестових даних, які складають 20% від повного набору.

Таблиця 1

Значення відхилень для вибраних методів

Метод	Середньоквадратичне відхилення
Лінійна регресія	89 086
Дерева рішень	31 199
Випадковий ліс	19 829

Одним із способів перевірки є використання коефіцієнту детермінація. Це коефіцієнт (R^2), який вимірює, наскільки добре статистична модель прогнозує результат. Результат представлений залежною змінною моделі. Найнижче можливе значення R^2 дорівнює 0, а найвище – 1. Простіше кажучи, чим краще модель надає прогнози, тим ближче її R^2 буде до 1 [5]. Як бачимо на рис. 5, для нашої моделі цей коефіцієнт складає 0,87, що є гарним показником.

```
X_test = strat_test_set.drop("price", axis=1)
Y_test = strat_test_set["price"].copy()
X_test_prepared = full_pipeline.transform(X_test)
forest_reg.score(X_test_prepared, Y_test)
```

✓ 0.4s

```
0.8715301849235442
```

Рис. 5 Коефіцієнт детермінація

Висновки та перспективи подальших пошуків у напрямі дослідження.

Розроблена модель створює перспективи для ефективного аналізу та прогнозування цін на житло, а також може слугувати важливим інструментом для різних груп зацікавлених сторін, включаючи покупців, продавців, інвесторів та різноманітні фінансові установи.

Важливо враховувати необхідність постійного вдосконалення та оновлення моделей, особливо у зв'язку зі змінами в економічному та соціокультурному середовищі. Зазначаю, що успішність моделі значно залежить від якості та достовірності початкових даних, а отже, систематична перевірка та корекція інформації є невід'ємною частиною впровадження таких рішень.

Це дослідження підкреслює потенціал машинного навчання для удосконалення процесів прийняття рішень на ринку нерухомості та може послужити важливим кроком у напрямку створення ефективних та точних інструментів для аналізу та прогнозу цінових тенденцій.

Список літератури

1. Асаул А.М., Павлов В.І., Пилипенко І.І., Павліха Н.В., Кривов’язюк І.В. Економіка нерухомості : навч. посіб., 2-ге вид.– К:, Кондор, 2006. 336 с.
2. Манцевич Ю. М. Житло: проблеми та перспективи. Київ : Профі, 2004. 360 с.
3. Григорків В. С., Ярошенко О. І., Філіпчук Н. В. Нейронні мережі та їх використання для прогнозування тенденцій ринку нерухомості. Науковий вісник НЛТУ України. 2012. Вип. 22.5. С. 324-330.
4. Воронін В. О., Мамчин М. М., Лянце Е. В. Прогнозне моделювання тенденцій розвитку ринку нерухомості. Вісник Національного університету "Львівська політехніка". 2012. № 735 : Логістика. С. 38-46.
5. Coefficient of Determination (R^2) | Calculation & Interpretation. Scribbr. URL: https://www.scribbr.com/statistics/coefficient-of-determination (дата звернення: 15.11.2023 р.).

6. Manil Wagle. Predicting House Prices using Machine Learning. Medium. URL: https://medium.com/@manilwagle/predicting-house-prices-using-machine-learning-cab0b82cd3f (дата звернення: 15.11.2023 р.).