

ІНФОРМАЦІЙНА СИСТЕМА ПРОГНОЗУВАННЯ ВАРТОСТІ КОМЕРЦІЙНИХ ТА ЖИТЛОВИХ ПРИМІЩЕНЬ НА РИНКУ НЕРУХОМОСТІ

Вінницький національний технічний університет

Анотація

Представлено розробку програмного модуля для автоматизованого визначення вартості комерційної та житлової нерухомості на основі методу машинного навчання Random Forest. Random Forest забезпечує детальний аналіз великого обсягу інформації про історичні дані про нерухомість, включаючи характеристики об'єктів, макроекономічні показники та локальну інфраструктуру. Проведено попередню обробку даних та виявлення аномалій методом Isolation Forest, в процесі розробки здійснено порівняння точності моделі та візуалізацію ключових компонентів, що впливають на ціноутворення.

Ключові слова: нерухомість, прогнозування вартості, Random Forest, машинне навчання, аналіз даних, Isolation Forest.

Abstract

The development of a software module for the automated valuation of commercial and residential real estate based on the Random Forest machine learning method is presented. Random Forest enables a detailed analysis of a large volume of historical real estate data, including property characteristics, macroeconomic indicators, and local infrastructure. Data preprocessing and anomaly detection were performed using the Isolation Forest method. During the development process, model accuracy was compared, and key components influencing pricing were visualized.

Keywords: real estate, price prediction, Random Forest, machine learning, data analysis, Isolation Forest.

Вступ

Важливою складовою сучасної світової економіки є ринок нерухомості, відповідно ключовим питанням для покупців, продавців, ріелторів та інших учасників ринку є питання точного прогнозування вартості об'єктів. З огляду на динамічні зміни в ціні та велику кількість факторів впливу на ринок, актуальним є впровадження автоматизованих методів оцінки нерухомості. Стандартні методи визначення вартості наразі є недоцільними для використання через спотворення прогнозів в залежності від величини вибірки даних для оцінки, та в залежності від кількості критеріїв, які впливають на формування ціни [1].

Постійний зріст об'єму інформації в світі вимагає більш детального її аналізу, розділення за категоріями та відсіювання даних, які спотворюють дослідження. Саме тому доцільним є застосування методів машинного навчання для аналізу та прогнозування вартості приміщень нерухомості, зокрема використання алгоритму Random Forest. Перевагою цього методу ансамблевого навчання є висока точність прогнозів, стійкість до перенавчання та здатність працювати з різними типами багатовимірних даних [2, 3].

Метою роботи є розробка методу для визначення вартості об'єктів на ринку нерухомості на основі алгоритму машинного навчання Random Forest.

Результати дослідження

Алгоритм Random Forest – це ансамблевий метод машинного навчання на основі дерев

рішень. Він створює множину незалежних дерев під час навчання, де кожне дерево тренується на випадковій підмножині даних (через бутстрепінг) та використовує лише частину ознак при розбитті вузлів. Така випадковість зменшує кореляцію між деревами, знижує ризик перенавчання та підвищує точність прогнозування. Метод Random Forest успішно обробляє великі набори інформації завдяки можливості паралельних обчислень, що прискорює навчання моделі. Особлива перевага методу – стійкість до перенавчання через усереднення результатів багатьох незалежних дерев, що критично для аналізу ринку нерухомості з його зашумленими або неповними даними. Крім того даний метод добре обробляє викиди (нетипові ціни на об'єкти) та ефективно працює з категоріальними змінними (район, тип будівлі, якість ремонту) без додаткової обробки, на відміну від інших алгоритмів [3]. Метод забезпечує стабільні оцінки важливості різних характеристик нерухомості, що допомагає краще розуміти який фактор має більший вплив на формування ціни. На рисунку 1 наведено схему роботи алгоритму Random Forest.

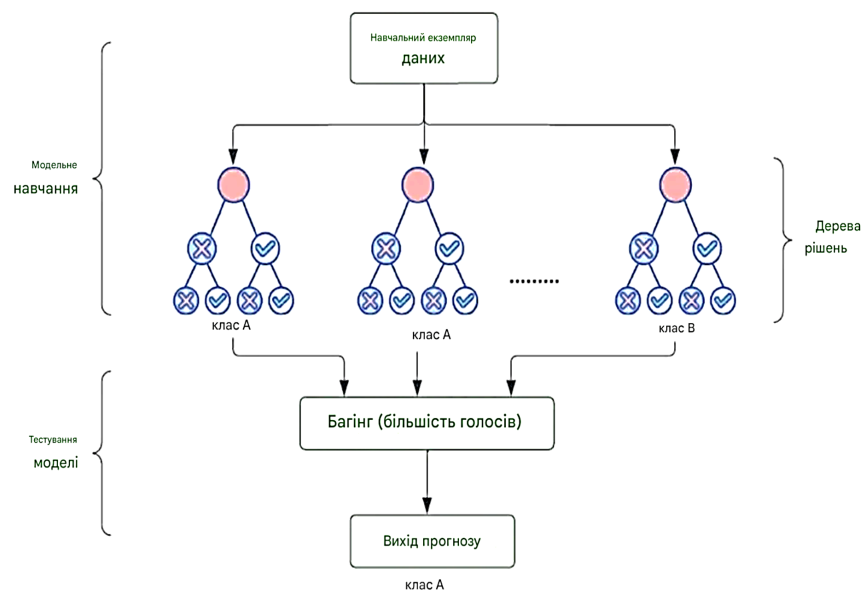


Рисунок 1 – Схема роботи алгоритму Random Forest [1, 2]

Для виявлення аномалій у вибірці даних, що аналізується, використовується метод Isolation Forest. Для кожної точки алгоритм визначає типівість її властивостей для наявної вибірки. Наприклад, об'єкт з дуже низькою ціною при великій площі або з високою ціною при малій площі може класифікуватися як аномалія. Алгоритм Isolation Forest базується на методі побудови ізольованих дерев для виявлення аномалій [3]. Його ключова ідея: аномалії легше ізолювати від основних даних, оскільки вони розташовані віддалено або в зонах із низькою щільністю. Спочатку алгоритм створює підмножини даних і для кожної будує дерево, що рекурсивно розділяє інформацію за випадковими ознаками та пороговими значеннями. Глибина кожного вузла $h(x)$ показує кількість розділень, необхідних для ізоляції точки. Значення глибини вузла $h(x)$ у дереві для точки x визначається кількістю поділів, що встановлюють її положення. Середня глибина дерева позначається як $c(n)$, де n – загальна кількість точок у дереві, і становить середнє значення для дерев з n об'єктами. Після побудови дерев алгоритм оцінює аномальний бал для кожної точки x :

$$S(x, n) = 2^{-\frac{E(h(x))}{c(n)}},$$

де $E(h(x))$ – середня глибина вузла для точки x в усіх деревах, а $c(n)$ – нормалізатор для очікуваної глибини в дереві з n об'єктами [4].

Значення $s(x, n)$ варіюється від 0 до 1:

- $s(x, n) \approx 1$ – точка є аномалією (легко ізолюється, глибина

мала);

- $s(x, n) \approx 0$ – точка є нормальною (важко ізолюється, глибина велика).

Під час дослідження було проаналізовано набір даних ціноутворення нерухомості. Масив критеріїв містить як багатовимірні дані, так і заздалегідь ідентифіковані аномалії. Основними факторами, які аналізувались при дослідженні на вплив вартості комерційних та житлових приміщень були: показники ціни за приміщення [35000...150 000] USD; площа приміщення в м²; рік побудови; кількість кімнат; віддаленість приміщення від центру міста; тип будівлі (квартира, будинок, комерційний, новобудова); наявність інфраструктури (школи, лікарні, торговельні центри поблизу тощо).

Проведене дослідження показало, що на ефективність роботи методу впливає кількість дерев, яка використовується при навчанні моделі. Для візуалізації роботи моделі в програмній реалізації розроблено графік порівняння прогнозованих і реальних цін, наведений на рисунку 2. Блакитні точки тестових даних демонструють щільне групування навколо лінії прогнозованих цін, особливо в середньому ціновому діапазоні (100-120 тисяч доларів), що підтверджує хорошу точність моделі в цьому сегменті. Для зміни чутливості роботи моделі можливо змінити кількість дерев, що використовуються. Для оцінки зміни точності використовуються метрики - Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) та Коефіцієнт детермінації (R^2) [5].

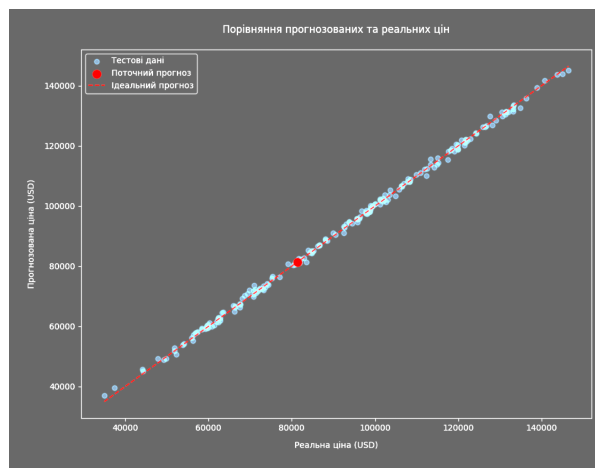


Рис. 2. Порівняльна діаграма прогнозованих та реальних цін

Для покращення результатів точності моделі виконувалося збільшення кількості дерев з 100 од. до 250 од. при прогнозуванні ціни на нерухомість. Аналіз метрик якості показав, що при таких змінах модель демонструє середню абсолютну помилку (MAE) близько 2 тис. дол. та середньоквадратичне відхилення (RMSE) приблизно 2,6 тис. дол.. Ці значення свідчать про високу точність роботи моделі, особливо враховуючи широкий ціновий діапазон та різноманіття об'єктів нерухомості, які оцінювались. Дані критерії та візуалізація за допомогою графіків надає можливість виконувати планування усім учасникам ринку нерухомості автоматизовано.

Висновки

Визначено, що запропонований підхід із використанням методу Random Forest дозволяє досягти високої точності прогнозування вартості об'єктів нерухомості з середньою абсолютною помилкою близько 2 тис. дол. У порівнянні із стандартними методами даний алгоритм є більш продуктивним та доцільним для використання при роботі з великою кількістю даних. Даний метод дозволяє усім учасникам ринку виконувати швидко та ефективно прогнозування вартості комерційних та житлових приміщень на ринку нерухомості, при аналізі великої кількості критеріїв.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Random Forest Algorithm in Machine Learning – [Електронний ресурс] – Режим доступу: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>
2. Класифікація дерев рішень – [Електронний ресурс] – Режим доступу: <https://towardsdatascience.com/decision-tree-classifier-explained-a-visual-guide-with-code-examples-for-beginners-7c863f06a71e>.
3. What is Isolation Forest? – [Електронний ресурс] – Режим доступу: <https://www.geeksforgeeks.org/what-is-isolation-forest/>
4. Isolation Forest – [Електронний ресурс] – Режим доступу: <https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.IsolationForest.html>.
5. Метрики бібліотеки sc-learn – [Електронний ресурс] – Режим доступу: https://scikit-learn.org/dev/modules/generated/sklearn.metrics.r2_score.html.

Примак Анастасія Вікторівна – студент групи КН-21Б, факультет інформаційних інтелектуальних технологій та автоматизації, Вінницький технічний національний університет, Вінниця.

Іванчук Ярослав Володимирович – д-р техн. наук, професор кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Шептяков Ігор Олександрович – аспірант кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Prymak Anastasiia V. – Department of Information Intellectual Technologies and Automation, Vinnytsia National Technical University, Vinnytsia.

Ivanchuk Yaroslav V. – Dr. Sc. (Eng.), Professor, Professor of the Department of Computer Sciences, Vinnytsia National Technical University, Vinnytsia.

Sheptiakov Ihor A. – postgraduate degree, Department of Computer Sciences, Vinnytsia National Technical University, Vinnytsia.