https://doi.org/10.31649/mccs2022.16

УДК 004.8:338+332.628

А. Р. Богачук[1]

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПЕРЕДБАЧЕННЯ ЦІНИ ПРОДАЖУ БУДИНКІВ У КІНГ-КАУНТІ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

[1]Вінницький національний технічний університет

Продаж та купівля нерухомості, зокрема житла, будинків є надзвичайно важливими для нашого життя. Більшість людей звертаються в агентства нерухомості до ріелторів з метою придбання якісного житла та водночас по оптимально вигідній для покупця ціні. Варто покладатись не лише на особисту оцінку або оцінку сторонніх експертів, але також використовувати системи передбачення ціни, які за допомогою ознак будинку (площа, кількість поверхів, місце розташування, кількість спальних кімнат, рік побудови тощо) спроможні передбачати можливу його ціну. Стаття присвячена задачі по підвищенні точності передбачення ціни продажу будинків у Кінг-Каунті методами машинного навчання шляхом створення інформаційної технології передбачення цієї ціни. Здійснено аналіз продажу та купівлі нерухомості, попередньо запропоновано ознаки, які мають вплив на ціноутворення будинків. Виконано вибір датасету та опис його основних ознак, попереднє очищення даних, проведено розвідувальний аналіз даних, запропоновано правило фільтрації аномальних значень для обраного датасету, вибрано множину можливих моделей, здійснено їх тренування та вибрано оптимальну серед них, наведено та проаналізовано результат роботи моделей, для порівняння наведено точність передбачення аналогічних рішень. Отримано оптимальну регресійну модель LGBM, її застосування дозволило отримати точність передбачення 0.876, що є більшим за 0.82, як у найкращого аналога. Наукова новизна полягає у тому, що дістала подальший розвиток інформаційна технологія передбачення ціни продажу будинків у Кінг-Каунті з використанням методів машинного навчання, яка дозволяє підвищити точність такого передбачення у порівнянні з аналогами.

Ключові слова: інформаційна технологія, розвідувальний аналіз даних, передбачення ціни, будинок, ознаки, моделі машинного навчання.

Вступ

Сьогодні ми можемо спостерігати велику кількість будинків, які продаються. Деякі з них перебувають у процесі зведення, інші вже введені в експлуатацію. При цьому вартість квадратних метрів в об’єктах значно відрізняється. Ціни залежать від типу будівлі, міста, району, техніки будівництва, площі, планування, стану та багатьох інших факторів.

У зв’язку з цим виникає питання складання вірної ціни продажу на будинок. Дане питання дуже актуальне у наш час, та напевне не менш актуальним залишатиметься у найближчі роки, а можливо й у майбутньому.

Оцінка будинку – послуга, без якої не обійтись у багатьох випадках. Фактично будь-які операції з нерухомим майном вимагають розрахунку його ринкової вартості [1].

Необхідність визначення того, скільки коштує будинок, потрібно в різних життєвих ситуаціях. Наприклад [1]:

– якщо оформляється спадщина, визначається реальна вартість будинку. Адже податкові зобов’язання за об’єкт, що успадковується, лягають на плечі спадкоємця. Занижена вартість також не схвалюється. У податківців можуть виникнути додаткові питання, що спричинить виплату додаткових податкових відрахувань;
– при внесенні власності до статутного капіталу підприємства;
– якщо будинку завдано шкоди;
– власник хоче застрахувати житло;
– оцінка вартості будинку під час розлучення;
– угоди купівлі-продажу тощо.

Якщо ж йдеться про котеджне село або містечко, як, наприклад Кінг Каунті (округ штату Вашингтон, США), то на оцінку впливатимуть внутрішня інфраструктура, віддаленість від центру, стан екології, наявність прибудинкової ділянки, охорони на території та навіть забудовник. Переваги за будь-якими пунктами підвищують цінність об'єкта, отже – і його вартість.

Інформаційна технологія передбачення ціни продажу будинків складається з розв'язання таких задач:

– вибір оптимальних інформаційних технологій;
– вибір датасету, огляд основних ознак та попереднє очищення даних;
– проведення розвідувального аналізу даних;
– вибір оптимальної моделі, створення інформаційної технології та її застосування для передбачення даних.

Метою статті є підвищення точності передбачення ціни продажу будинків у Кінг-Каунті методами машинного навчання шляхом створення інформаційної технології передбачення цієї ціни.

Вибір датасету, огляд основних ознак та попереднє очищення даних

Для проведення дослідження використовуватимемо дані США, Кінг-Каунті (по 21613 будинках) із датасету «House Sales in King County, USA» на базі платформи, тобто без обмежень на копіювання і використання [2]. Для реалізації обрані програмні пакети та бібліотеки мови програмування Python.

Дані містять такі ознаки (рис. 1) [2]:

– дата, коли будинок був розпроданий ("date");
– ціна будинку ("price");
– кількість спалень у будинку ("bedrooms");
– кількість ванних кімнат ("bathrooms");
– площа будинку у квадратних футах ("sqft_living ");
– площа земельної ділянки ("sqft_lot");
– кількість поверхів будинку ("floors");
– чи є вид на набережну ("waterfront ");
– чи переглядали будинок ("view");
– стан будинку за шкалою від 1 до 5 ("condition");
– загальна оцінка, на основі системи класифікації графства Кінг за шкалою від 1 до 11 ("grade");
– площа будинку не враховуючи підвальне приміщення ("sqft_above");
– площа підвального приміщення будинку ("sqft_basement");
– рік побудови ("yr_built");
– поштовий індекс будинку ("zipcode");
– координати розташування будинку, широта та довгота ("lat", "long");
– площа житлового приміщення найближчих 15 сусідів ("sqft_living15");
– площа земельних ділянок найближчих 15 сусідів ("sqft_lot15").

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_b
0	221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180	0
1	538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170	400
2	180000.0	2	1.00	770	10000	1.0	0	0	3	6	770	0
3	604000.0	4	3.00	1960	5000	1.0	0	0	5	7	1050	910
4	510000.0	3	2.00	1680	8080	1.0	0	0	3	8	1680	0

Рис. 1. Приклад ознак будинків, що містяться у датасеті

Проведемо попереднє очищення даних. Отримавши інформацію ознак датасету виявлено, що є ознаки, які мають велику кількість нульових значень, або не несуть ніякої цінності при передбаченні, тому їх видалено, це такі ознаки, як: "zipcode", "view", "waterfront", "yr_renovated".

Після очищення даних отримано датасет з 15 ознаками по 21609 будинках.

Розвідувальний аналіз даних

Застосуємо метод Describe для наступних значень квантилів (1%, 5%, 10%, 50%, 90%, 92%, 97%, 99%), наведено на рис. 2.

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	condition	gr
count	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	2
mean	5.400881e+05	3.370842	2.114757	2079.899736	1.510697e+04	1.446213	3.409430	7.
std	3.671272e+05	0.930062	0.770163	918.440897	4.142051e+04	0.551894	0.650743	1.
min	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	1.000000	1.
1%	1.535004e+05	2.000000	1.000000	720.000000	1.013120e+03	1.000000	3.000000	5.
5%	2.100000e+05	2.000000	1.000000	940.000000	1.800000e+03	1.000000	3.000000	6.
10%	2.450000e+05	2.000000	1.000000	1090.000000	3.322200e+03	1.000000	3.000000	6.
50%	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.000000	3.000000	7.
90%	8.870000e+05	4.000000	3.000000	3250.000000	2.139760e+04	2.000000	4.000000	9.
92%	9.500000e+05	5.000000	3.250000	3420.000000	2.851660e+04	2.000000	4.000000	9.
93%	9.980000e+05	5.000000	3.250000	3510.000000	3.484832e+04	2.000000	5.000000	1(
94%	1.063560e+06	5.000000	3.250000	3630.000000	3.768116e+04	2.000000	5.000000	1(
96%	1.259040e+06	5.000000	3.500000	3920.000000	5.065816e+04	2.000000	5.000000	1(
97%	1.388000e+06	5.000000	3.500000	4140.000000	6.743684e+04	2.000000	5.000000	1(
99%	1.964400e+06	6.000000	4.250000	4978.800000	2.130080e+05	3.000000	5.000000	1
max	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.000000	5.000000	1:

Рис. 2. Значення квантилів для ключових ознак будинків

З рисунка 2 видно, що суттєво відрізняється ціна будинків з квантилем 94% та 96% і мінімальним квантилем та 5%, аналогічно ціні будинку проаналізовано інші ознаки за квантилями. Це дозволило визначити межі для фільтрування аномальних значень, подане у вигляді коду на Python, рис. 3.

```
train0 = train0[(
                (train0['price'] <= 1000000) &
                (train0['price'] > 170000) &
                (train0['bathrooms'] <= 4) &
                (train0['condition'] > 2.5) &
                (train0['grade'] != 4) &
                (train0['sqft_lot15'] > 1300) &
                (train0['sqft_lot15'] < 44000) &
                (train0['sqft_lot'] > 1500) &
                (train0['sqft_lot'] < 70000) &
                (train0['sqft_living'] > 700) &
                (train0['yr_built'] > 1925) &
                (train0['bedrooms'] > 0) &
                (train0['bedrooms'] < 7)
                )]
```

Рис. 3. Приклад коду на Python застосування фільтрів за верхньою та нижньою межею значень по ряду ознак

За правилом з рисунка 3 виконано фільтрування даних, яке ще зменшило датасет до 15825 будинків, гістограма для яких наведена на рисунку 4.

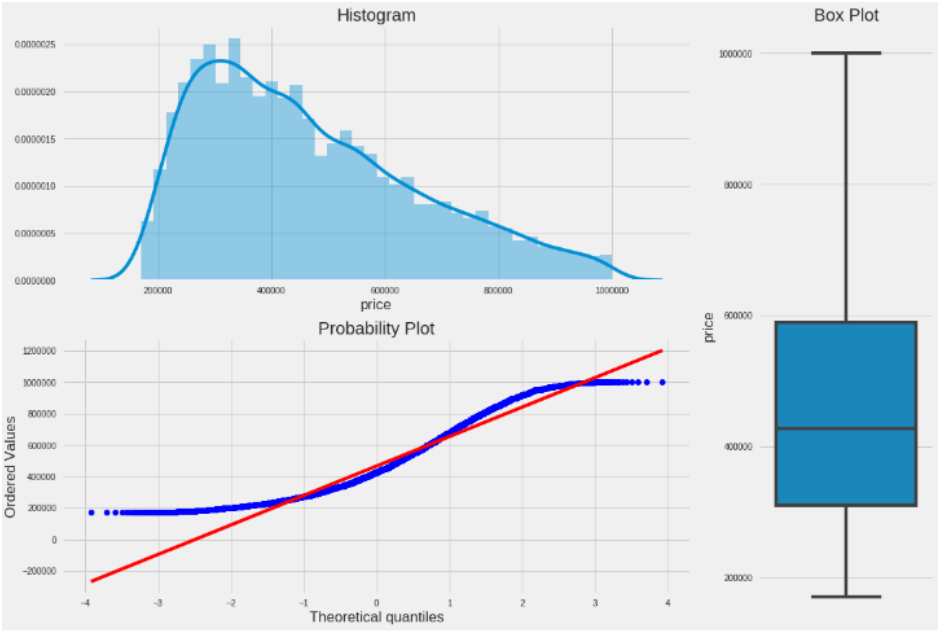

Рис. 4. Перевірка на аномальні дані методами Matplotlib, Pandas та Seaborn

З рисунка 4 видно гістограму розподілення даних за ціною будинків, більша частина будинків має ціну від 200 тис. доларів до 450 тис. доларів, будинки з ціною вищою за 800 тис. доларів трапляються рідше, підтвердженням того є діаграма «коробка з вусами» з неї теж випливає, що середня ціна будинку варіюється між значеннями 300 тис. доларів та 600 тис. доларів.

За допомогою моделей лінійної регресії, XGBoost та LGB побудовано діаграму важливості ознак (рис. 5), з якої видно, що на ціну будинку найбільший вплив мають такі ознаки, як: місце розташування будинку, площа земельної ділянки, площа будинку, площа земельних ділянок та будинків найближчих сусідів, рік побудови.

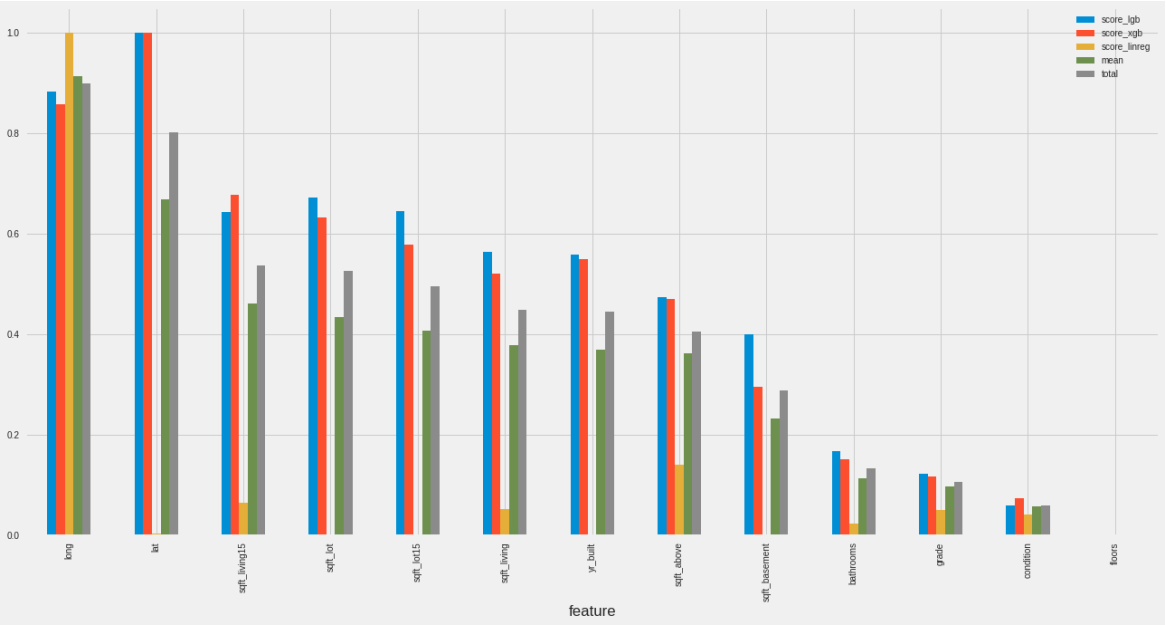

Рис. 5. Діаграма важливості ознак

Вибір і навчання моделей та їх застосування для передбачення даних

Задача передбачення ціни на будинки належить до виду машинного навчання з учителем (контрольоване навчання). А одним із найкращих варіантів розв'язання даної задачі є розв'язок за допомогою моделей регресії та моделей, які побудовані на основі дерев рішень [3].

Для визначення ціни будинку необхідно дослідити залежність ціни від ознак того чи іншого будинку. Пропонується застосувати моделі-регресори (RandomForestRegressor, XGBRegressor, LightGBM, BaggingRegressor, ExtraTreesRegressor, LinearRegression, MLPRegressor) [4].

Їх застосування дозволило ранжувати дані за точністю R^2-критерію (рис. 6).

	Model	r2_train	r2_test	d_train	d_test	rmse_train	rmse_test
2	LGBM	94.42	87.60	6.77	10.04	4,527,283.13	6,847,700.24
1	XGB	95.71	86.99	6.21	10.26	3,971,028.02	7,014,609.60
0	Random Forest	97.00	84.81	4.65	11.16	3,320,970.87	7,581,097.88
3	BaggingRegressor	97.14	84.56	4.56	11.16	3,241,689.74	7,641,576.41
4	ExtraTreesRegressor	99.82	83.67	0.12	11.42	805,340.86	7,858,618.51
6	MLPRegressor	69.20	68.61	17.01	17.33	10,635,722.46	10,896,497.60
5	Linear Regression	67.68	67.29	17.73	17.92	10,894,560.31	11,122,762.64

Рис. 6. Ранжування за R^2-критерієм результатів передбачення моделей

Аналіз показав, що найкращою моделлю за R^2-критерієм є модель LGBM, її застосування дозволило отримати точність передбачення 0.876.

Дане рішення є кращим за точністю від аналогів, які використовують подібні моделі, таких, як:

– «Predicting House Prices». $R^2 = 0.82$ [5];
– «House Price Predictions (R^2 0.82)» [6].

Висновки

Дослідження набору даних, що містить інформацію про продажі будинків США (Кінг-Каунті) показало, що для точного передбачення ціни потрібно провести розгорнутий розвідувальний аналіз даних, відфільтрувати помилкові та аномальні дані, відкинути недоцільні ознаки. Побудовано діаграму важливості ознак. Наступним кроком можна переходити до тренування моделей та порівняння їх точності, для визначення оптимальної. Визначено, що для розв'язання задачі передбачення ціни доцільно обрати моделі-регресори. Вибрано та натреновано 7 моделей. Оптимальною визначено модель LGBM, її застосування дозволило отримати точність передбачення 0.876, що є більшим за 0.82, як у найкращого аналога. Здійснено підвищення точності передбачення ціни продажу будинків у Кінг-Каунті методами машинного навчання шляхом створення інформаційної технології передбачення цієї ціни.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1]. *Проведення оцінювання будинку.* [Electronic resource]. Available: https://pareto.com.ua/ua/blog/yak-provoditsya-ocinka-budinku/

[2]. *House Sales in King County, USA*. [Electronic resource]. Available: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

[3]. К. Ю. Кононова, *Машинне навчання: методи та моделі*. Харків: ХНУ імені В. Н. Каразіна, 2020, 301 с.

[4]. *Supervised Learning API Overview*. [Electronic resource]. Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[5]. *Predicting House Prices.* [Electronic resource]. Available: https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices

[6]. *House Price Predictions (R^2 0.82).* [Electronic resource]. Available: https://www.kaggle.com/code/rotemgb/house-price-predictions-r-2-0-82

REFERENCES

[1]. *Conducting a home appraisal*. [Electronic resource]. Available: https://pareto.com.ua/ua/blog/yak-provoditsya-ocinka-budinku/

[2]. *House Sales in King County, USA*. [Electronic resource]. Available: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

[3]. K. Y. Kononova, *Machine learning: methods and models*. Kharkiv: V. N. Karazin Kharkiv National University, 2020, 301 c.

[4]. *Supervised Learning API Overview*. [Electronic resource]. Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[5]. *Predicting House Prices.* [Electronic resource]. Available: https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices

[6]. *House Price Predictions (R^2 0.82).* [Electronic resource]. Available: https://www.kaggle.com/code/rotemgb/house-price-predictions-r-2-0-82

Богачук Андрій Русланович — студент кафедри системного аналізу та інформаційних технологій, e-mail: fkca.2ict.bar@gmail.com.

Вінницький національний технічний університет

A. R. Bohachuk[1]

Information Technology Predicting the Sale Prices of Houses in King-County by Machine Learning Methods

[1]Vinnytsia National Technical University

The sale and purchase of the real estate, in particular housing, and houses are extremely important for our life. Most people turn to real estate agencies to realtors in order to purchase quality housing and at the same time at the best price for the buyer. You should rely not only on personal assessment or assessment of third-party experts but also use price prediction systems that, using the features of the house (area, number of floors, location, number of bedrooms, year of construction, etc.), are able to predict its possible price. The report is devoted to the task of improving the accuracy of predicting the sale price of houses in King-County using machine learning methods by creating an information technology for predicting this price. The analysis of sales and purchases of real estate has been carried out, and signs that have an impact on the pricing of houses have been previously proposed. The dataset was selected, and its main features were described, preliminary data cleaning was carried out, exploratory data analysis was carried out, a rule for filtering anomalous values for the selected dataset was proposed, many possible models were selected, they were trained, and the optimal one was selected, the result of the models was presented and analyzed, for comparison the prediction accuracy of similar solutions is given.

An optimal LGBM regression model was obtained, and its application made it possible to obtain a prediction accuracy of 0.876, which is more than 0.82, as in the best analog. The scientific novelty lies in the fact that the information technology for predicting the sale price of houses in King-County has been further developed using machine learning methods, which makes it possible to increase the accuracy of such a prediction compared to analogs.

Keywords: information technology, intelligence analysis of data, price prediction, house, feature, machine learning models.

Bohachuk Andrii R. — student of the Department of System Analysis and Information Technologies, e-mail: fkca.2ict.bar@gmail.com.