УДК 004.8:338+332.628

О.С. Дідик

Ю.М. Паночишин

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПРОГНОЗУВАННЯ ЦІН НА ПРИВАТНІ ЖИТЛОВІ БУДИНКИ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Запропоновано інформаційну технологію прогнозування ціни продажу будинків методами машинного навчання та описані основні етапи розв'язання задачі.

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

Abstract

Information technology for predicting the sale price of houses using machine learning methods is proposed and the main stages of problem solving are described.

Keywords: information technology, exploratory data analysis, price prediction, house, features, machine learning models.

Вступ

Купівля власного житла є однією з найважливіших фінансових інвестицій у житті людини. Це не лише приносить емоційне задоволення, але й є практичним кроком. Власний дім забезпечує простір для життя і роботи, а також дарує сім'ї відчуття стабільності та захищеності.

Однак, знайти ідеальний будинок може бути складним завданням, а знайти ідеальний будинок за хорошою ціною – ще більше. Велика кількість факторів можуть впливати на ціну будинку, таких як розташування, розмір, вік, стиль та рівень оновлення. Тому, розуміння того, як ці фактори впливають на вартість будинку, може бути корисним при придбанні будь-якої нерухомості.

Окрім того, знання ринку нерухомості та здатність передбачати ціни на будинки може бути цінним не тільки для приватних осіб, але й для компаній та інвесторів. Вони можуть використовувати дані, щоб зрозуміти ринок нерухомості та ризики, пов'язані з інвестуванням у нерухомість [1].

Мета і задачі дослідження. Метою даного дослідження є підвищення точності прогнозування цін на приватні житлові будинки за допомогою методів машинного навчання шляхом розробки та впровадження інформаційної технології, що базується на аналізі даних ринку нерухомості.

Для досягнення цієї мети необхідно вирішити такі завдання:

– провести всебічний аналіз об'єкта дослідження, визначивши ключові фактори, що впливають на ціну;

– виконати розвідувальний аналіз даних для виявлення патернів та аномалій, а також здійснити обробку вхідних даних для забезпечення якості прогнозування;

– розробити та протестувати моделі машинного навчання для прогнозування цін на основі зібраних даних.

Результати проведеного дослідження

Для проведення дослідження використано дані, що описують (майже) кожен аспект житлових будинків у місті Еймс, штат Айова із датасету «House Prices - Advanced Regression Techniques» на базі платформи Kaggle [2]. Для реалізації інформаційної технології були обрані програмні пакети та бібліотеки мови програмування Python.

Дані містять такі ознаки:

– MSSubClass – клас будівлі;
– MSZoning – класифікація району;
– LotFrontage – лінійні метри вулиці, що з'єднана з об'єктом нерухомості;
– LotArea – розмір ділянки в квадратних футах;
– Street – вулиця, тип дорожнього доступу;
– Alley – тип алеї;
– LotShape – загальна форма ділянки;
– LandContour – рівність ділянки;
– Utilities – тип наявних інженерних комунікацій;
– LotConfig – конфігурація ділянки;
– LandSlope – нахил ділянки;
– Neighborhood – райони в межах міста Еймс;
– Condition1 – близькість до головної дороги або залізниці;
– Condition2 – близькість до головної дороги або залізниці (за наявності);
– BldgType – тип житла;
– HouseStyle – кількість поверхів;
– OverallQual – загальна якість матеріалів та оздоблення;
– OverallCond – загальна оцінка стану будинку;
– YearBuilt – рік побудови;
– YearRemodAdd – дата реконструкції;
– RoofStyle – тип даху;

Приклад тренувального набору даних (train) показано на рисунку 1.

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	...	PoolArea	Po
0	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	Na
1	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	...	0	Na
2	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	...	0	Na
3	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	...	0	Na
4	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	...	0	Na
5	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	Inside	...	0	Na
6	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	Na
7	60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	Corner	...	0	Na
8	50	RM	51.0	6120	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	Na
9	190	RL	50.0	7420	Pave	NaN	Reg	Lvl	AllPub	Corner	...	0	Na

Рисунок 1 – Тренувальний набір даних

Після перегляду дата сету було виявлено багато пропущених значень, тому для точного прогнозу було проведено розвідувальний аналіз даних. Для перевірки розподілу відносно змінної SalePrice (Ціна продажу) ознак були побудовані графіки залежності категоріальних змінних (рис. 2).

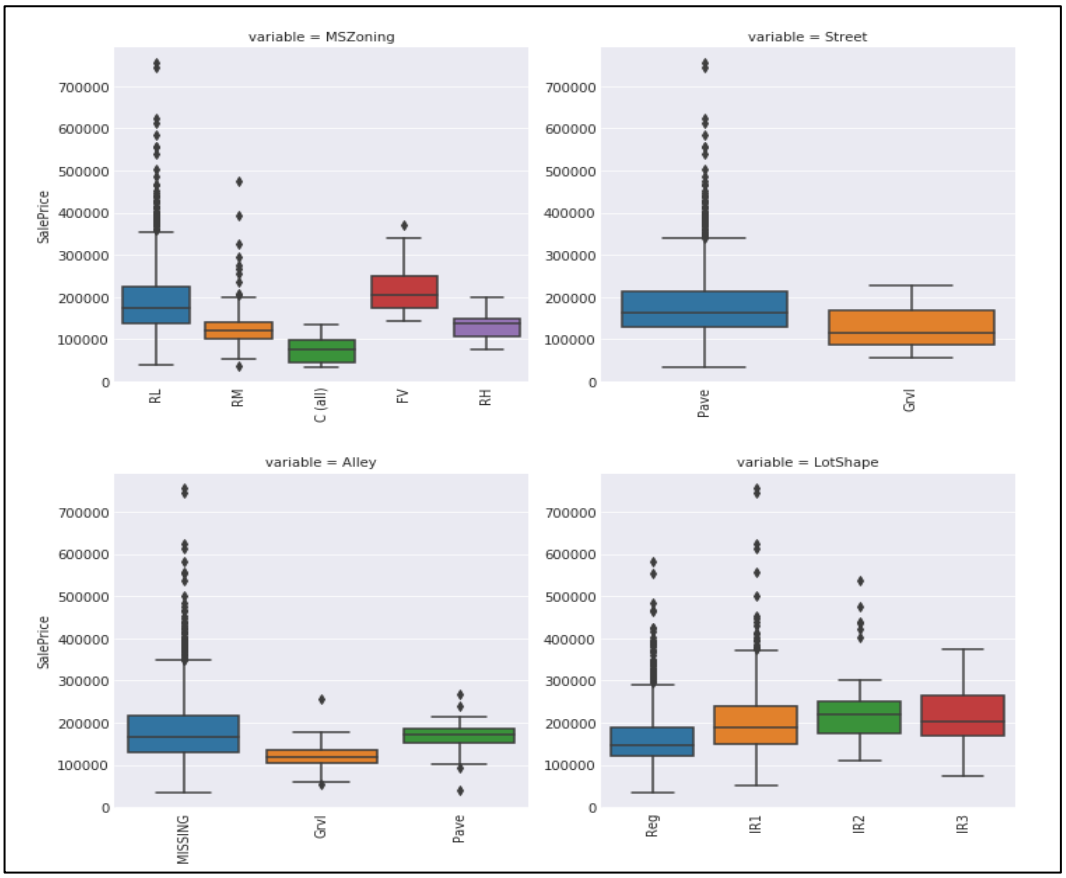

Рисунок 2 – Графіки залежності категоріальних змінних відносно SalePrice

Виконано перевірку вхідних даних на наявність відхилення значень цільової змінної. Після перевірки було видалено значення, які мали значне відхилення і могли впливати на точність моделі (рис. 3).

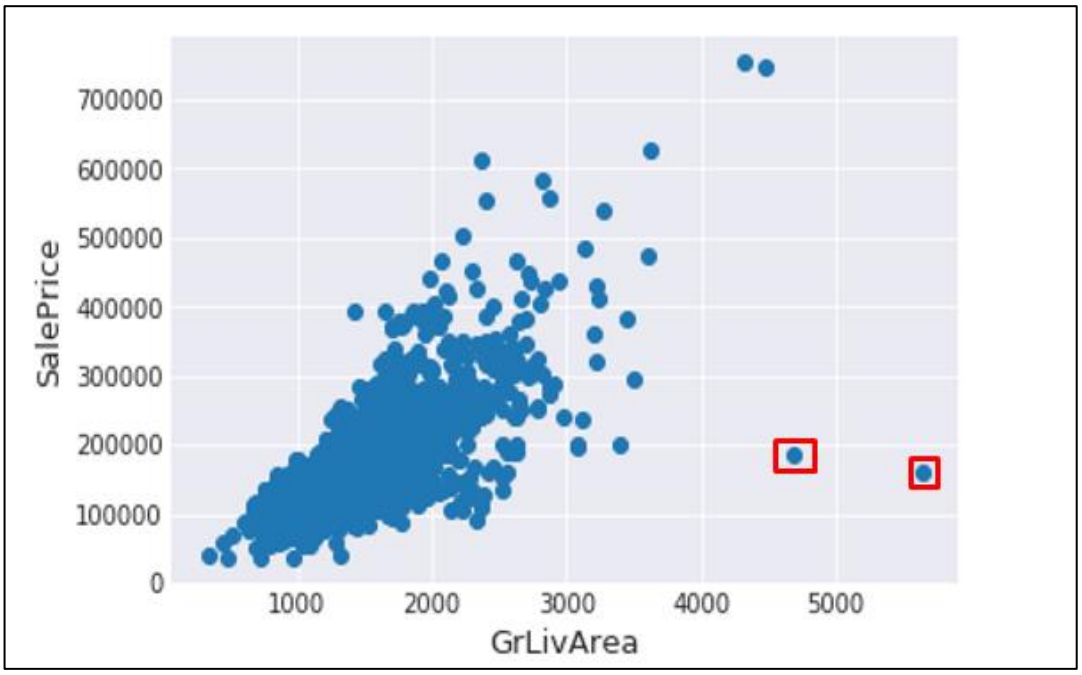

Рисунок 3 – Перевірка відхилень

Для покращення передбачення моделей виконано логарифмічне перетворення даних, після якого цільова змінна підпорядковувалась нормальному розподілу. Також застосовано перетворення Бокса-Кокса, яке використовується для зменшення викривленості (асиметрії) в розподілі даних.

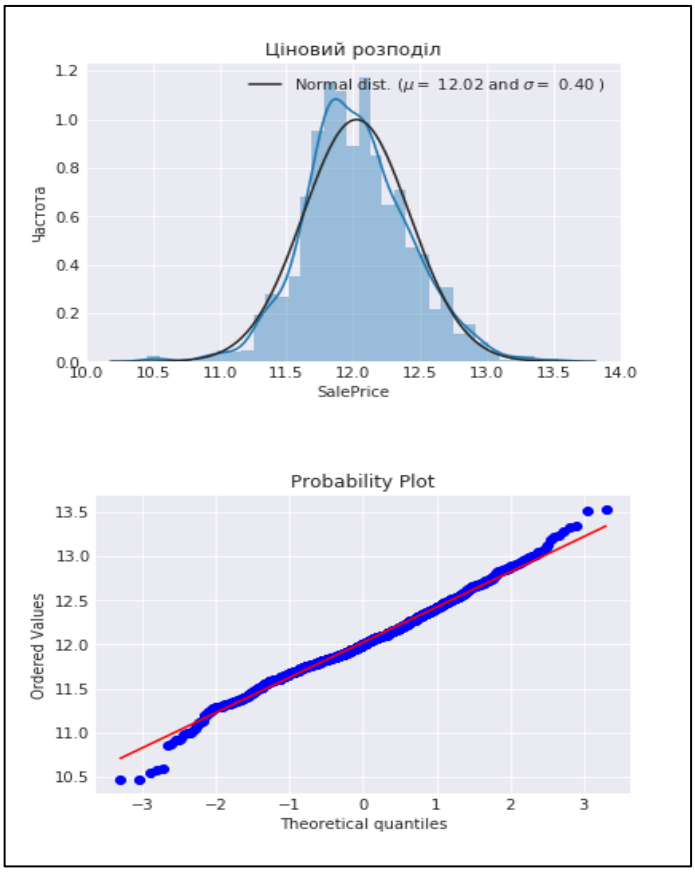

Рисунок 4 – Розподіл цін та цільова змінна після логарифмічного перетворення

Для здійснення прогнозу було взято моделі Enet, KRR, Gboost, XGBoost, LightGBM [3]. Виконана перехресна перевірка середньоквадратичної помилки (RMSE) для вказаних моделей. Далі було прийнято рішення виконати стекування моделей для покращення результату передбачення. Для цього виконано усереднення моделей Enet, KRR та Gboost, а lasso була використана як метамодель [4].

За результати тестування найкращий результат передбачення за метрикою RMSE показла модель LightGBM що на 34% краще ніж у аналога (табл. 1).

Таблиця 1 – Порівняння результатів точності прогнозування розробленої програми із аналогом.

Власні результати		Результати аналога	
Назва моделі	Похибка за метрикою RMSE	Назва моделі	Похибка за метрикою RMSE
LightGBM	0.073	LightGBM	0.119
XGBoost	0.080	XGBoost	0.114
StackedRegressor	0.074	Lasso	0.112

Висновки

Дослідження набору даних, що містить інформацію про продаж будинків у США (штат Айова), показало, що для точного прогнозування цін необхідно провести детальний розвідувальний аналіз даних, відфільтрувати помилкові та аномальні значення, а також виключити недоцільні ознаки. Наступним кроком було навчання моделей та порівняння їхньої точності для визначення оптимальної. Встановлено, що для розв'язання задачі прогнозування цін на приватні житлові будинки доцільно використовувати модель LightGBM, оскільки вона продемонструвала найкращий результат.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Що визначає і впливає на ціну нерухомості. [Електронний ресурс]. URL: https://novebti.ua/ua/blog/chto-opredelyaet-i-vliyaet-na-cenu-nedvizhimosti/
2. Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. [Електронний ресурс]. URL: https://kaggle.com/competitions/house-prices-advanced-regression-techniques
3. Vijay Kanade, What Is Machine Learning? Definition, Types, Applications, and Trends for 2022. URL: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/
4. Олександр Дідик. Kagle Notebook «Prediction», «House Prices - Advanced Regression Techniques» [Електронний ресурс]. URL: https://www.kaggle.com/code/olexandrdidyk/prediction

Дідик Олександр Сергійович– студент групи 2КН-23м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail; olexadrdidyk23@gmail.com

Паночишин Юрій Миколайович **-** доцент кафедри комп'ютерних наук, кандидат технічних наук, Вінницький національний технічний університет, м. Вінниця. email: y.panochyshyn@vntu.edu.ua

Didyk Oleksandr S. – student of Intelligent Information Technologies and Automation Department, Vinnytsia National Technical University, Vinnytsia, email; olexadrdidyk23@gmail.com

Panochyshyn Yury M. - Associate Professor of the Department of Computer Sciences, Candidate of Technical Sciences, Vinnytsia National Technical University, Vinnytsia. email: y.panochyshyn@vntu.edu.ua