МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Чорноморський національний університет імені Петра Могили

Факультет комп'ютерних наук

Кафедра інтелектуальних інформаційних систем

ДОПУЩЕНО ДО ЗАХИСТУ

Завідувач кафедри інтелектуальних інформаційних систем

_____Юрій КОНДРАТЕНКО

«____»_____ 2024 р.

КВАЛІФІКАЦІЙНА РОБОТА

НА ЗДОБУТТЯ ОСВІТНЬОГО СТУПЕНЯ МАГІСТРА

ІНФОРМАЦІЙНА СИСТЕМА ПРОГНОЗУВАННЯ ЦІН НА РИНКУ НЕРУХОМОСТІ

Спеціальність 124 Системний аналіз

Освітня програма «Системний аналіз»

Здобувач _____ Яна ПОТУЖНЯ

«___»_____ 2024 р.

Керівник д-р техн. наук, доцент _____ Ірина КАЛІНІНА

«___»_____ 2024 р.

Миколаїв – 2024

Чорноморський національний університет імені Петра Могили
(повне найменування закладу вищої освіти)

Факультет	Комп'ютерних наук
Кафедра	Інтелектуальних інформаційних систем
Рівень вищої освіти	Другий (магістерський)
Освітній ступень	Магістр
Спеціальність	124 Системний аналіз
Освітня програма	Системний аналіз

ЗАТВЕРДЖУЮ

Завідувач кафедри інтелектуальних інформаційних систем

_____ Юрій КОНДРАТЕНКО

«____» _____ 2024 р.

ЗАВДАННЯ
на кваліфікаційну роботу здобувача

Потужньої Яни Олександрівни
(прізвище, ім'я, по батькові здобувача)

1. Тема кваліфікаційної роботи: «Інформаційна система прогнозування цін на ринку нерухомості».

Керівник роботи: Калініна Ірина Олександрівна, в. о. професора кафедри ІІС, д-р техн. наук, доцент.

Затверджена наказом ЧНУ ім. Петра Могили від «03» червня 2024 р. № 140/1.

2. Строк представлення кваліфікаційної роботи «17» грудня 2024 р.

3. Очікуваний результат роботи та початкові дані, якщо такі потрібні: розроблений прототип інформаційної системи прогнозування цін на ринку нерухомості; проаналізований теоретичний матеріал, поставлена задача створення інформаційної системи прогнозування цін на ринку нерухомості.

4. Перелік питань, що підлягають розробці: аналіз сучасних інформаційних систем та методів прогнозування; аналіз загальної структури інформаційної системи прогнозування цін на ринку нерухомості; реалізація інформаційної системи прогнозування цін на ринку нерухомості; порівняльний аналіз отриманих результатів для визначення найбільш ефективних методів прогнозування.
5. Перелік графічних матеріалів: презентація.

Керівник роботи _____ Ірина КАЛІНІНА
(Особистий підпис) *(Власне ім'я ПРІЗВИЩЕ)*

Здобувач _____ Яна ПОТУЖНЯ
(Особистий підпис) *(Власне ім'я ПРІЗВИЩЕ)*

Дата видачі завдання «07» червня 2024 р.

КАЛЕНДАРНИЙ ПЛАН
кваліфікаційної роботи

Тема: Інформаційна система прогнозування цін на ринку нерухомості

№	Найменування роботи	Початок	Закінчення	Примітки
1	Отримання завдання на виконання КР	03.06.2024	07.06.2024	Виконано
2	Аналіз предметної області та постановка задачі	10.06.2024	20.06.2024	Виконано
3	Огляд літературних джерел за темою кваліфікаційної роботи, зокрема аналіз публікацій та аналогічних систем, щодо прогнозування цін на ринку нерухомості	21.06.2024	01.07.2024	Виконано
4	Побудова структури інформаційної системи та вибір методів для вирішення поставленої задачі	01.09.2024	25.10.2024	Виконано
5	Реалізація інформаційної системи з аналізом отриманих результатів	26.10.2024	21.11.2024	Виконано
6	Перший попередній захист КР на засіданні комісії кафедри	22.11.2024	22.11.2024	Виконано
7	Корегування роботи за результатами попереднього захисту	23.11.2024	05.12.2024	Виконано
8	Другий попередній захист КР на засіданні комісії кафедри	06.12.2024	06.12.2024	Виконано
9	Доробка та остаточне оформлення КР	07.12.2024	10.12.2024	Виконано
10	Подання КР, її електронної копії та інших документів (відгуку, рецензії) до захисту	16.12.2024	17.12.2024	Виконано

Керівник роботи _____ Ірина КАЛІНІНА
(Особистий підпис) *(Власне ім'я ПРІЗВИЩЕ)*

Здобувач _____ Яна ПОТУЖНЯ
(Особистий підпис) *(Власне ім'я ПРІЗВИЩЕ)*

Дата складання календарного плану
«19» червня 2024 р.

АНОТАЦІЯ

до кваліфікаційної роботи
здобувачки групи 607м ЧНУ ім. Петра Могили

Потужньої Яни Олександрівни

на тему: **“ІНФОРМАЦІЙНА СИСТЕМА ПРОГНОЗУВАННЯ ЦІН НА РИНКУ НЕРУХОМОСТІ”**

Актуальність даного дослідження полягає в необхідності точного прогнозування цін на ринку нерухомості для підтримки прийняття обґрунтованих рішень в умовах високої динамічності цього ринку. Використання сучасних методів машинного навчання дозволяє значно підвищити точність прогнозів і автоматизувати процес оцінки вартості нерухомості. Розробка програмного забезпечення для цієї задачі дасть можливість за допомогою інтерактивного інтерфейсу здійснювати швидкий і точний розрахунок вартості об'єктів на основі ключових факторів, що впливають на ціноутворення. Це, в свою чергу, дозволить знизити ризики при прийнятті рішень на ринку нерухомості та підвищити ефективність інвестиційних стратегій.

Об’єктом дослідження є процес прогнозування цін на ринку нерухомості за допомогою інформаційної системи, що аналізує різноманітні фактори, які впливають на вартість нерухомості.

Предметом дослідження є інформаційна система, що здійснює прогнозування цін на ринку нерухомості. Вона включає в себе методи та алгоритми аналізу даних, які використовуються для обробки інформації про об'єкти нерухомості та фактори, що впливають на формування цін.

Метою дослідження є прогнозування цін на ринку нерухомості з використанням інформаційної системи, яка здійснює аналіз різноманітних факторів, що впливають на вартість об'єктів.

В результаті виконання роботи було досліджено кілька методів для прогнозування цін на ринку нерухомості, зокрема лінійну регресію, множинну регресію, поліноміальну регресію, дерева рішень, випадковий ліс та XGBoost.

Проаналізовано ефективність кожного з методів на основі тестових даних та оцінено точність моделей. Визначено основні переваги та недоліки кожного методу в контексті прогнозування цін на нерухомість. Також було розроблено інформаційну систему, в якій реалізовано ці методи для автоматизованого прогнозування цін на основі введених користувачем даних.

Дана робота складається з чотирьох розділів. Кожен розділ відповідно присвячений: аналізу предметної області; структурі інформаційної системи, моделям і методам, що використані у роботі; аналізу та попередній обробці даних; реалізації інформаційної системи прогнозування цін на ринку нерухомості та аналізу отриманих результатів. Загальний обсяг роботи – 113 сторінок. Кваліфікаційна робота містить 9 додатків, 55 рисунків, 8 таблиць і 45 джерел посилання.

Ключові слова: прогнозування цін , ринок нерухомості, лінійна регресія, множинна регресія, XGBoost, випадковий ліс, поліноміальна регресія, дерева рішень, інформаційна система.

ABSTRACT

to the qualification work by the student of the group 607m of Petro Mohyla Black Sea National University

Potuzhnia Yana

“ INFORMATION SYSTEM FOR FORECASTING PRICES ON THE REAL ESTATE MARKET ”

A relevance of this study lies in the need for accurate price prediction on the real estate market to support informed decision-making in the context of the market's high dynamics. The use of modern machine learning methods significantly enhances the accuracy of predictions and automates the property valuation process. The development of software for this task will enable quick and precise property value calculations through an interactive interface, based on key factors that influence price formation. This, in turn, will help reduce risks in decision-making on the real estate market and improve the effectiveness of investment strategies.

An object of research is the process of price prediction in the real estate market using an information system that analyzes various factors influencing property value.

A subject of the research is an information system that performs price prediction in the real estate market. It includes methods and algorithms for data analysis used to process information about property objects and factors influencing price formation.

A purpose of the study is to predict prices in the real estate market using an information system that analyzes various factors influencing the value of properties.

As a result of the work, several methods for predicting prices in the real estate market were investigated, including linear regression, multiple regression, polynomial regression, decision trees, random forest, and XGBoost. The effectiveness of each method was analyzed based on test data, and the accuracy of the models was evaluated. The main advantages and disadvantages of each method in the context of real estate price prediction were identified. Additionally, an information system was developed that implements these methods for automated price prediction based on user-entered data.

This work consists of four sections. Each of them is devoted to: the analysis of the subject area; the structure of the information system, models, and methods used in the work; data analysis and preprocessing; the implementation of the real estate price prediction information system, and the analysis of the obtained results.

The overall scope of the work is 113 pages. Thesis contains 9 applications, 55 figures, 8 tables and 45 references in it.

Key words: price prediction, real estate market, linear regression, multiple regression, XGBoost, random forest, polynomial regression, decision trees, information system.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

GBR (англ. Gradient Boosting Regressor)	регресор градієнтного бустингу
XGBoost (англ. Extreme Gradient Boosting)	екстремальний градієнтний бустинг
LGBM (англ. Light Gradient Boosting Machine)	легкий градієнтний бустинг
AIC (англ. Akaike Information Criterion)	критерій Акаїке
DW (англ. Durbin-Watson)	критерій Дарбіна-Уотсона
MSE (англ. Mean Squared Error)	середня квадратична помилка
F (англ. Fisher’s Statistic)	статистика Фішера
Theil (англ. Theil’s U Statistic)	коефіцієнт Тейла
MAPE (англ. Mean Absolute Percentage Error)	середня абсолютна помилка у відсотках

ВСТУП

У сучасному світі ринок нерухомості займає важливе місце в економіці, оскільки він не лише впливає на індивідуальні фінансові рішення, але й має суттєве значення для загального соціально-економічного розвитку країн. Ціни на нерухомість є результатом складної взаємодії багатьох факторів, зокрема економічних, соціальних, демографічних та політичних. Зростання цін на житло, зміни в попиті та пропозиції, а також вплив інфляційних процесів є лише частиною факторів, що впливають на ринок.

Ринок нерухомості є ключовим сегментом економіки, оскільки він впливає на безпосереднє життя населення та формує фінансові ресурси держави. Правильне прогнозування цін на нерухомість дозволяє оптимізувати інвестиційні рішення, сприяє стабільності ринку та допомагає уникати фінансових втрат для інвесторів і покупців.

В умовах глобалізації та інтеграції економік, ринок нерухомості стає все більш чутливим до змін у зовнішньому середовищі. Такі фактори, як економічні кризи, зміни в політиці, соціально-демографічні зміни, а також глобальні виклики, такі як пандемії, значно впливають на попит і пропозицію в цьому секторі.

Сучасні технології, такі як машинне навчання та аналіз даних, відкривають нові можливості для покращення точності прогнозів. Використання цих технологій дозволяє обробляти великі обсяги інформації та виявляти складні закономірності, що раніше були важкодоступними для традиційних статистичних методів. Це, у свою чергу, підвищує конкурентоспроможність на ринку та дозволяє учасникам ухвалювати більш обґрунтовані рішення на основі точних даних.

1 АНАЛІЗ РИНКУ НЕРУХОМОСТІ, ІСНУЮЧИХ МЕТОДІВ ТА ІНФОРМАЦІЙНИХ СИСТЕМ, ЩО ВИКОРИСТОВУЮТЬСЯ ДЛЯ ПРОГНОЗУВАННЯ ЦІН

1.1 Характеристика ринку нерухомості

Ринок нерухомості є складною і багатогранною системою економічних відносин, яка включає купівлю, продаж, оренду та інші форми обміну нерухомим майном. Він охоплює широкий спектр об'єктів, від житлових будинків і квартир до комерційних приміщень, офісних будівель, промислових об'єктів та земельних ділянок (див. рис. 1.1).

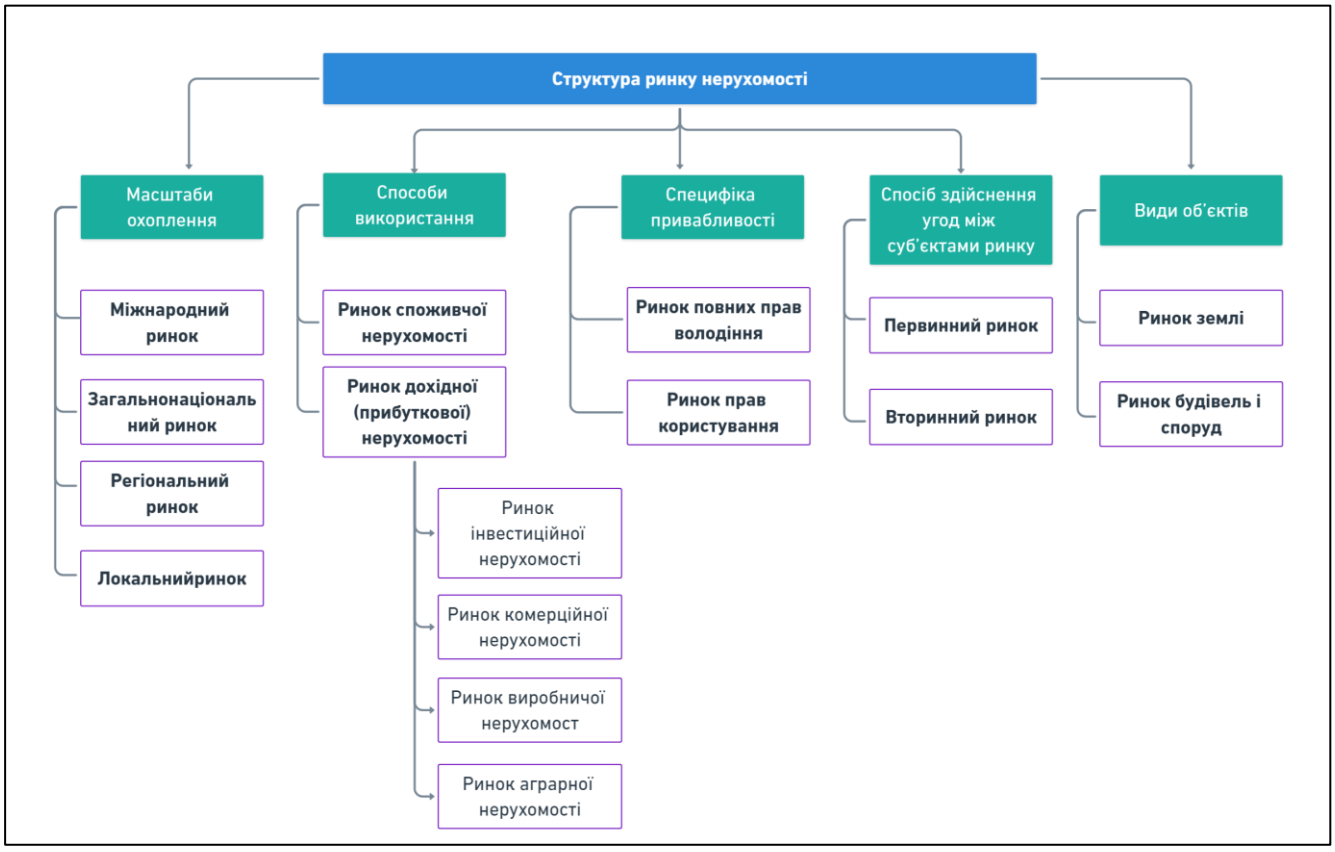

Рисунок 1.1 – Структура ринку нерухомості [1]

Кожен з цих сегментів має свої особливості, що вимагають специфічних підходів до оцінки вартості, аналізу попиту і пропозиції, а також розробки інвестиційних стратегій [1]. Ринок нерухомості відіграє важливу роль у формуванні

економічної стабільності, оскільки він тісно пов'язаний з банківським сектором через іпотечні кредити, фінансовими ринками через інвестиції та з державним управлінням через регуляторну політику. Крім того, ринок нерухомості є відображенням соціально-економічних процесів у суспільстві, таких як урбанізація, демографічні зміни, розвиток інфраструктури та технологічний прогрес. Він також впливає на рівень життя населення, забезпечуючи не тільки житлові потреби, але й створюючи робочі місця в будівельному секторі та суміжних галузях.

Первинний та вторинний ринки нерухомості представляють два основні сегменти ринку житлової нерухомості, які відрізняються за джерелом пропозиції та характером угод [2].

Первинний ринок нерухомості включає в себе новобудови, тобто об'єкти, що продаються вперше після їх будівництва. Це можуть бути нові житлові комплекси, котеджі, багатоквартирні будинки, які щойно зведені або ще будуються. Основною характеристикою первинного ринку є те, що нерухомість зазвичай нова, з сучасними будівельними технологіями і новими комунікаціями. Ціни на первинному ринку можуть бути нижчими на етапі будівництва і збільшуються після завершення об'єкта.

Вторинний ринок нерухомості охоплює продаж вже використовуваних або старих об'єктів нерухомості, які були в експлуатації і перепродаються їх власниками або агентствами нерухомості. Ціни на вторинному ринку можуть бути як нижчими, так і вищими за первинний ринок, залежно від розташування, стану об'єкта та ринкових умов.

Ринок нерухомості охоплює широкий спектр майна, що використовується для різних цілей, і кожен сегмент має свої специфічні характеристики, що визначають його динаміку та вплив на економіку.

Основні сегменти ринку нерухомості [2]:

– житлова нерухомість;

- комерційна нерухомість;
- промислова нерухомість;
- земельні ділянки;
- спеціалізована нерухомість.

Житлова нерухомість включає об'єкти, призначені для проживання людей. Це можуть бути квартири, приватні будинки, житлові комплекси та інше. Основні характеристики цього сегменту ринку - велика варіативність цін залежно від місця розташування, типу забудови та інфраструктури.

Комерційна нерухомість охоплює будівлі та приміщення, які використовуються для бізнес-цілей. Це можуть бути офісні будівлі, торгові центри, готелі, ресторани та інші комерційні об'єкти. Цей сегмент ринку характеризується високою динамікою цін, залежною від економічної активності, рівня розвитку бізнесу та інвестиційної привабливості регіону.

Промислова нерухомість включає заводи, фабрики, склади та інші об'єкти, які використовуються для виробництва та зберігання товарів. Цей сегмент ринку зазвичай має меншу варіативність цін, але велику залежність від економічної політики, рівня розвитку промисловості та логістики.

Земельні ділянки є ключовим компонентом ринку нерухомості, представляючи землю, яка може бути використана для різноманітних цілей, включаючи житлове, комерційне, промислове будівництво та сільське господарство. Вони відрізняються за призначенням, характеристиками та розташуванням, що впливає на їх вартість та використання.

Спеціалізована нерухомість представляє собою особливий сегмент ринку нерухомості, що включає об'єкти з конкретним призначенням, які важко або економічно не доцільно адаптувати для інших цілей. Цей тип нерухомості включає в себе ряд ключових категорій, таких як медичні установи, навчальні заклади, спортивні комплекси та культурні інституції.

Ці основні сегменти ринку нерухомості створюють різноманітні можливості для інвестицій, аналізу та прогнозування ринкових тенденцій. Кожен з них має свої специфічні особливості та фактори, що впливають на його розвиток і динаміку.

1.2 Фактори, що впливають на ціноутворення ринку нерухомості

Ціноутворення на ринку нерухомості є результатом взаємодії багатьох факторів, які впливають на вартість земельних ділянок, житлових та комерційних об'єктів. Розуміння цих факторів дозволяє краще аналізувати ринкові тенденції та приймати обґрунтовані рішення щодо інвестицій та купівлі нерухомості. Основними чинниками, що впливають на ціноутворення, є економічні, соціальні, політичні та демографічні [3].

Економічні умови мають безпосередній вплив на ринок нерухомості, що проявляється через кілька ключових аспектів.

Загальний рівень економічного зростання і валовий внутрішній продукт (ВВП) країни або регіону грають важливу роль у формуванні попиту на нерухомість і фінансову спроможність покупців. Коли економіка демонструє зростання, це зазвичай призводить до підвищення попиту на житлові та комерційні об'єкти, оскільки люди і компанії мають більші фінансові можливості для інвестицій у нерухомість. Це збільшення попиту може, в свою чергу, викликати зростання цін на ринку нерухомості.

Другим важливим чинником є інфляція, яка впливає на витрати на будівельні матеріали, працю та інші витрати, що безпосередньо впливають на ціну нових об'єктів нерухомості. Коли рівень інфляції зростає, це призводить до підвищення витрат на будівництво, що, відповідно, спричиняє зростання цін на нерухомість [3].

Не менш важливим є ринок праці, зокрема рівень зайнятості та доходів населення. Високий рівень зайнятості і стабільні доходи позитивно впливають на попит на житло, оскільки забезпечують фінансову спроможність населення для покупки або оренди нерухомості. Збільшення доходів і зменшення безробіття

сприяють підвищенню попиту на житлові об'єкти, що може призвести до зростання цін на нерухомість.

Соціальні фактори мають значний вплив на ціни на нерухомість через кілька ключових аспектів. Зміни в способі життя та уподобаннях населення можуть суттєво вплинути на попит на різні типи нерухомості. Наприклад, зростання популярності роботи на відстані може стимулювати підвищений попит на заміські або передміські житлові об'єкти, оскільки люди шукають більше простору і спокійніші умови для життя поза межами міських агломерацій [3].

Політична ситуація і державна політика мають значний вплив на ціноутворення на ринку нерухомості. Зміни в податковій політиці, такі як зміни в податках на нерухомість або на прибуток від продажу, можуть вплинути на інвестиційну привабливість ринку нерухомості. Зміни в податкових ставках або в умовах оподаткування можуть або стимулювати, або стримувати інвестиції в нерухомість, що, у свою чергу, вплине на ціни на ринку [3].

Ціноутворення на ринку нерухомості є складним процесом, що визначається взаємодією численних економічних, соціальних, політичних і демографічних чинників. Розуміння цих факторів допомагає краще прогнозувати зміни на ринку нерухомості, оцінювати інвестиційні можливості та адаптувати стратегії до поточних ринкових умов. Чіткий аналіз кожного з цих аспектів дозволяє краще орієнтуватися в ринкових тенденціях і приймати обґрунтовані рішення, що сприяють успішним інвестиціям та управлінню нерухомістю.

1.3 Аналіз ключових тенденцій на ринку нерухомості України

Ринок нерухомості в Україні характеризується значними коливаннями попиту та пропозиції, що відображаються у динаміці цін на житло в різних регіонах. Аналіз даних з різних міст України, дає змогу виявити ключові тенденції, що визначають ринок нерухомості.

Аналіз ключових тенденцій на ринку нерухомості Києва.

Київ, як столиця України, має один із найбільш активних ринків нерухомості в країні. Ціни на житло тут залишаються високими, відображаючи стабільний попит на житлову нерухомість. Середня вартість квадратного метра в Києві сягає **75 398 грн**. Високий попит зумовлений розвиненою інфраструктурою, наявністю бізнес-центрів та культурних об'єктів. Пропозиція на ринку нерухомості зростає завдяки активному будівництву нових житлових комплексів, що задовольняють різноманітні потреби покупців.

В Києві ціни на квартири значно варіюються в залежності від кількості кімнат. Згідно з даними сайту Dom.ria.com, середні ціни на квартири у Києві наступні [4] (див. табл. 1.1):

Таблиця 1.1 – Середні ціни на квартири у Києві

Тип квартири	**Середня ціна за м² (грн)**
Однокімнатні квартири	83 514
Двокімнатні квартири	87 076
Трикімнатні квартири	92 396
Чотирикімнатні квартири	99 384

Зі збільшенням кількості кімнат зростає й середня вартість квадратного метра житла. Це пов’язано з тим, що багатокімнатні квартири зазвичай мають більшу загальну площу і знаходяться в престижніших районах або нових житлових комплексах.

Ціни на квартири значно різняться між різними районами Києва, що пов'язано з розташуванням, наявністю інфраструктури, транспортною доступністю та престижністю району (див. рис. 1.2). У центральних районах, таких як Шевченківський і Печерський, середня вартість квадратного метра може досягати **102 900 – 123 480 грн** [4].

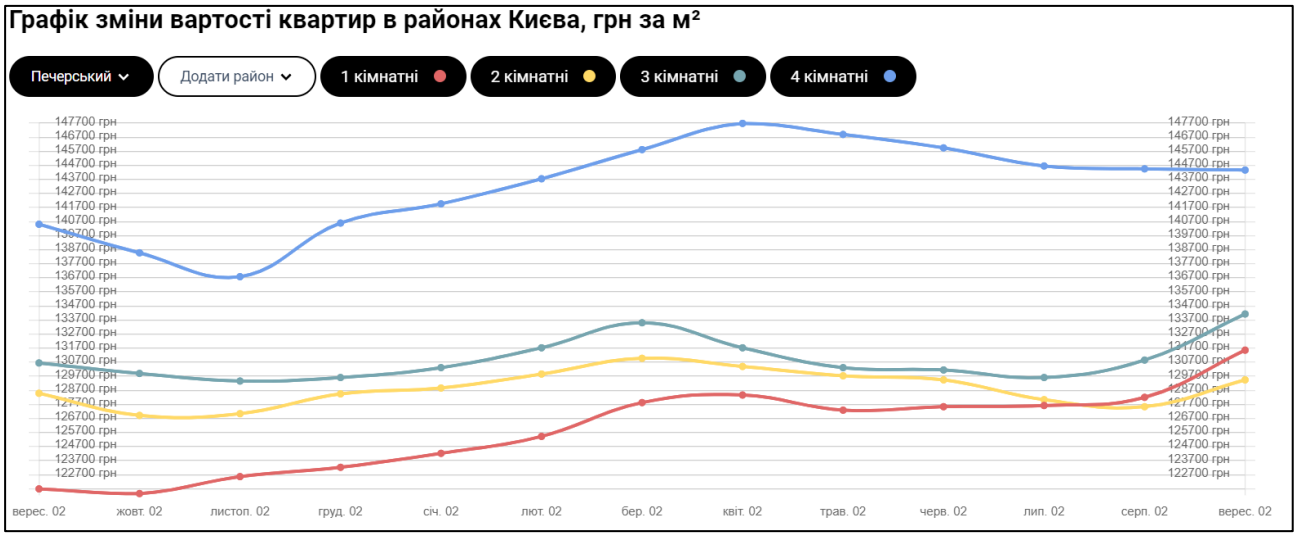

Рисунок 1.2 – Графік зміни вартості квартир в Печерському районі Києва

У віддалених районах, таких як Дарницький чи Святошинський, ціни нижчі – приблизно **61 740 – 74 088 грн** за квадратний метр (див. рис. 1.3).

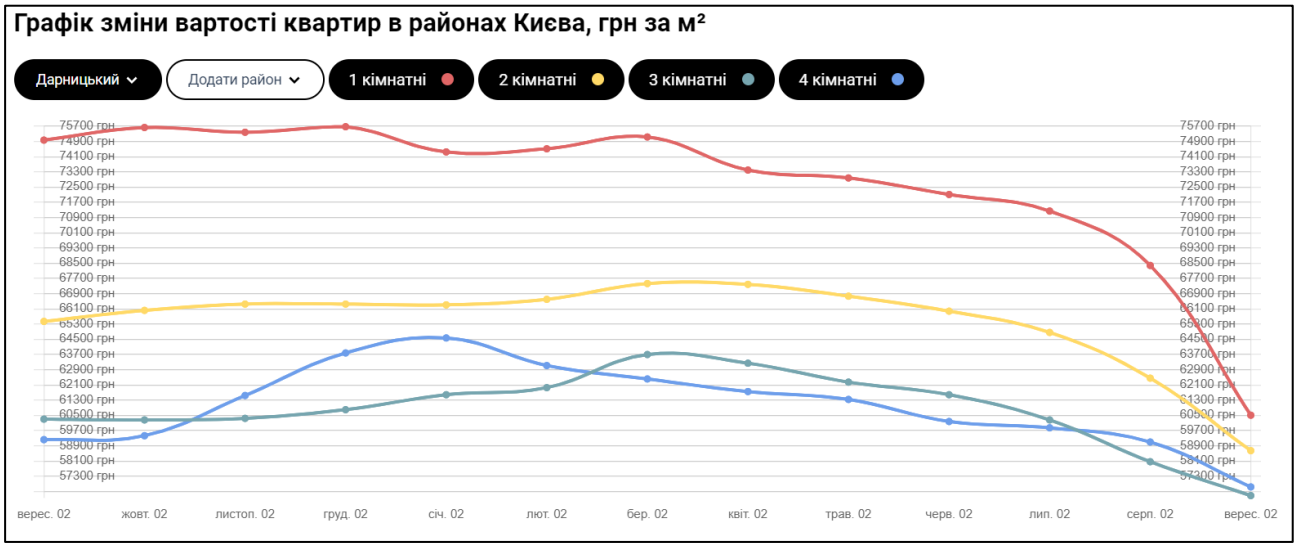

Рисунок 1.3 – Графік зміни вартості квартир в Дарницькому районі Києва

Відмінності в цінах на нерухомість у Києві пояснюються кількома факторами. Центральні райони мають розвинену інфраструктуру, що забезпечує зручний доступ до основних бізнес-центрів, культурних і розважальних об'єктів. Це робить їх більш привабливими для покупців. У свою чергу, віддалені райони пропонують житло за нижчими цінами, але також забезпечують доступ до

необхідної інфраструктури, хоча і на менш високому рівні. Тому ці райони часто вибирають ті, хто шукає доступніше житло або більшу площу за меншою ціною.

Аналіз ринку нерухомості в Києві показує, що динаміка попиту та пропозиції значно залежить від розташування та типу житла. Високі ціни в центрі міста зумовлені розвиненою інфраструктурою та престижністю районів, тоді як віддалені райони пропонують більш доступні варіанти для покупців. Пропозиція на ринку активно зростає завдяки будівництву нових житлових комплексів, що надає можливість задовольнити різноманітні потреби населення.

Аналіз ключових тенденцій на ринку нерухомості Одеси.

Одеса є одним з найбільших і найдинамічніших міст України, з важливим стратегічним значенням як для країни, так і для регіону в цілому. Це привабливе місто на узбережжі Чорного моря відоме своїм історичним і культурним значенням, що також відображається в його ринку нерухомості.

Ринок нерухомості в Одесі характеризується значним попитом на житло, зокрема, на квартири в центральних районах та біля узбережжя. Середня вартість квадратного метра в Одесі сягає **41 450 грн**.

Згідно з даними Dom.ria.com, ціни на квартири в Одесі варіюються залежно від кількості кімнат [5] (див. табл. 1.2):

Таблиця 1.2 – Середні ціни на квартири в Одесі

Тип квартири	Середня ціна за м² (грн)
Однокімнатні квартири	41 270
Двокімнатні квартири	45 833
Трикімнатні квартири	40 042
Чотирикімнатні квартири	35 274

Ціни на квартири значно різняться між різними районами Одеси, що пов'язано з їх розташуванням та рівнем розвитку інфраструктури. У Приморському районі середня вартість квадратного метра становить **41 572 грн**, що пояснюється близькістю до моря та розвиненою інфраструктурою, яка приваблює туристів та інвесторів (див. рис. 1.4).

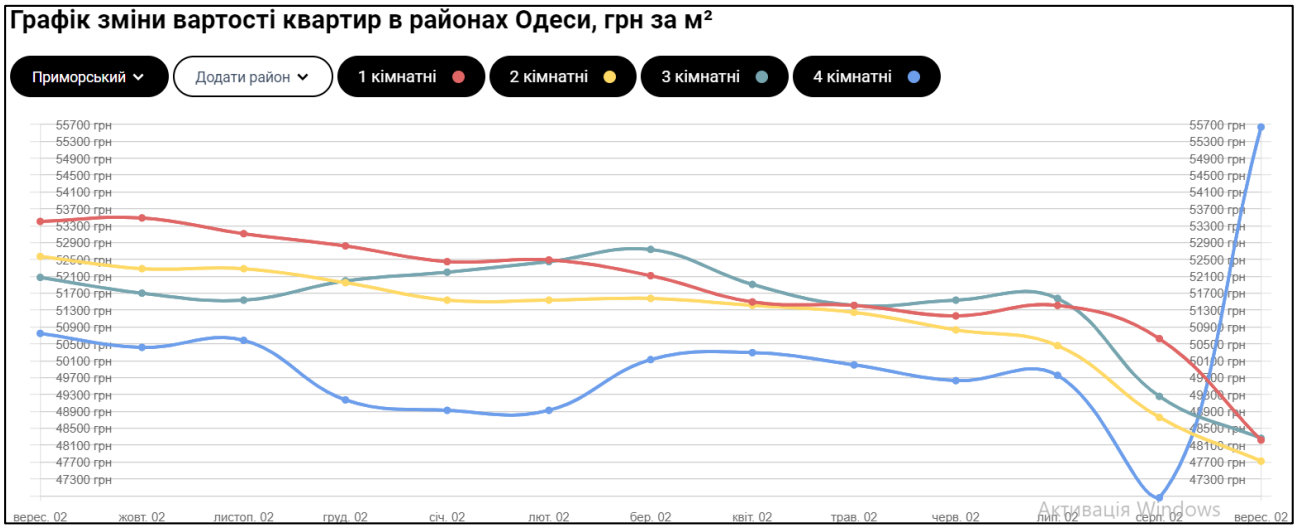

Рисунок 1.4 – Графік зміни вартості квартир в Приморському районі Одеси

У Київському районі ціна досягає **43 390 грн/м²** через стабільний попит і зручності для життя [5] (див. рис. 1.5).

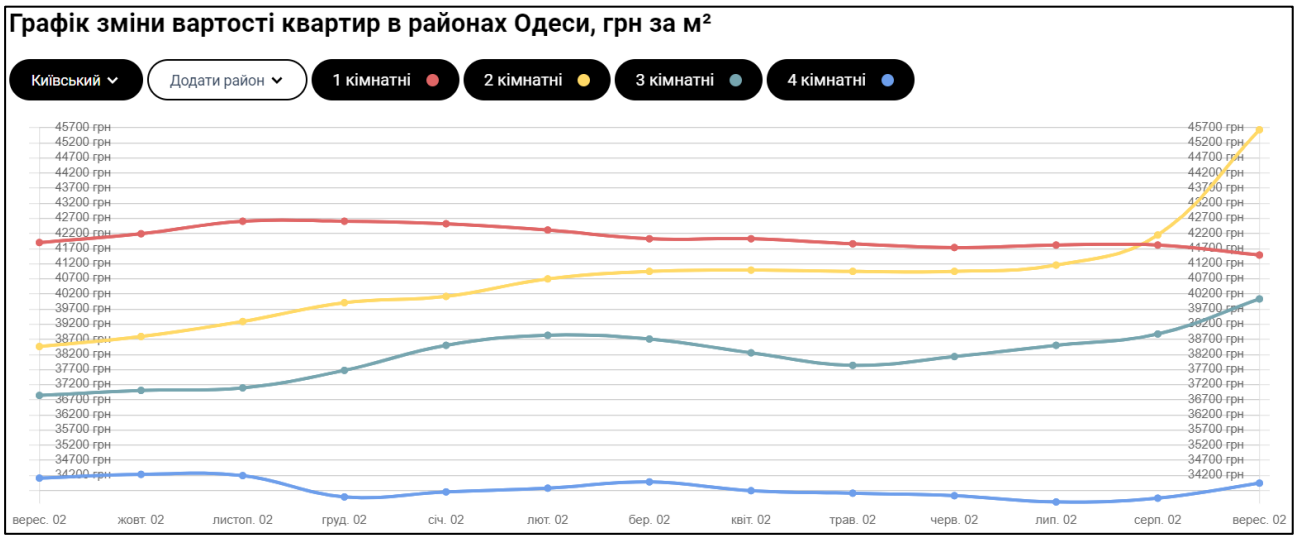

Рисунок 1.5 – Графік зміни вартості квартир в Київському районі Одеси

Великий Фонтан має найвищу ціну – **54 054 грн/м²**, що зумовлено престижністю району та новобудовами (див. рис. 1.6).

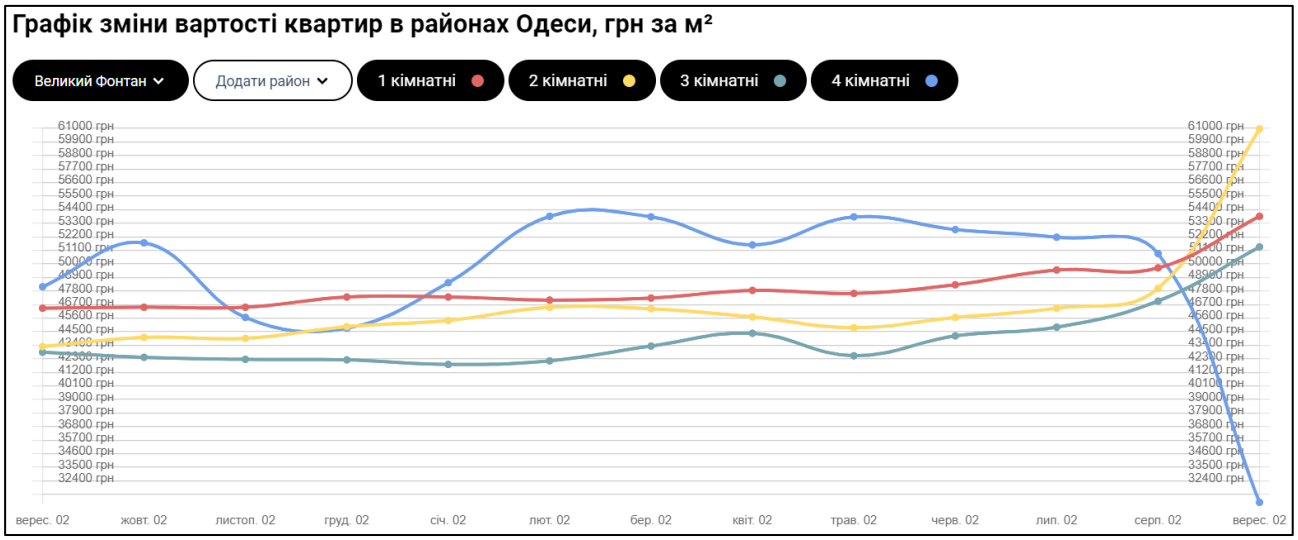

Рисунок 1.6 – Графік зміни вартості квартир в районі Великий Фонтан

Аналіз тенденцій на ринку житла в Одесі відзначає кілька ключових аспектів. По-перше, спостерігається стабільний попит на нерухомість у центральних районах, що зумовлено розвитком інфраструктури, наявністю культурних об'єктів та зручним доступом до моря. По-друге, зростає інтерес до новобудов, що відповідають сучасним стандартам, оскільки покупці шукають комфортні та енергоефективні рішення. Також, виявлено тенденцію до зміщення попиту в бік віддалених районів, де пропонуються більш доступні ціни, що робить їх привабливими для сімей, які шукають більшу площу за нижчу вартість. Таким чином, ринок житла в Одесі демонструє динамічний розвиток, враховуючи потреби різних категорій покупців.

Аналіз ключових тенденцій на ринку нерухомості Львова.

Львів, одне з найстаріших і культурно насичених міст України, відоме своєю архітектурою, історичними пам'ятками і культурними заходами. Це місто також є важливим економічним центром західної частини України.

Ринок нерухомості у Львові характеризується значними коливаннями попиту та пропозиції, що відображається у динаміці цін на житло в різних районах міста. Середня вартість квадратного метра у Львові сягає **60 890 грн**.

Згідно з даними сайту Dom.ria.com, ціни на квартири у Львові варіюються залежно від кількості кімнат [6] (див. табл. 1.3):

Таблиця 1.3 – Середні ціни на квартири у Львові

Тип квартири	**Середня ціна за м² (грн)**
Однокімнатні квартири	54 403
Двокімнатні квартири	59 226
Трикімнатні квартири	67 832
Чотирикімнатні квартири	80 699

Як і в інших великих містах, зі збільшенням кількості кімнат квартири у Львові збільшується середня вартість квадратного метра. Це пов'язано з тим, що багатокімнатні квартири часто розташовані в більш престижних районах або у нових житлових комплексах з покращеною інфраструктурою.

Середні ціни на квартири у Львові значно варіюються залежно від району. Наприклад, у центральних районах, таких як Галицький, ціна за квадратний метр може досягати близько **67 854 грн** (див. рис. 1.7).

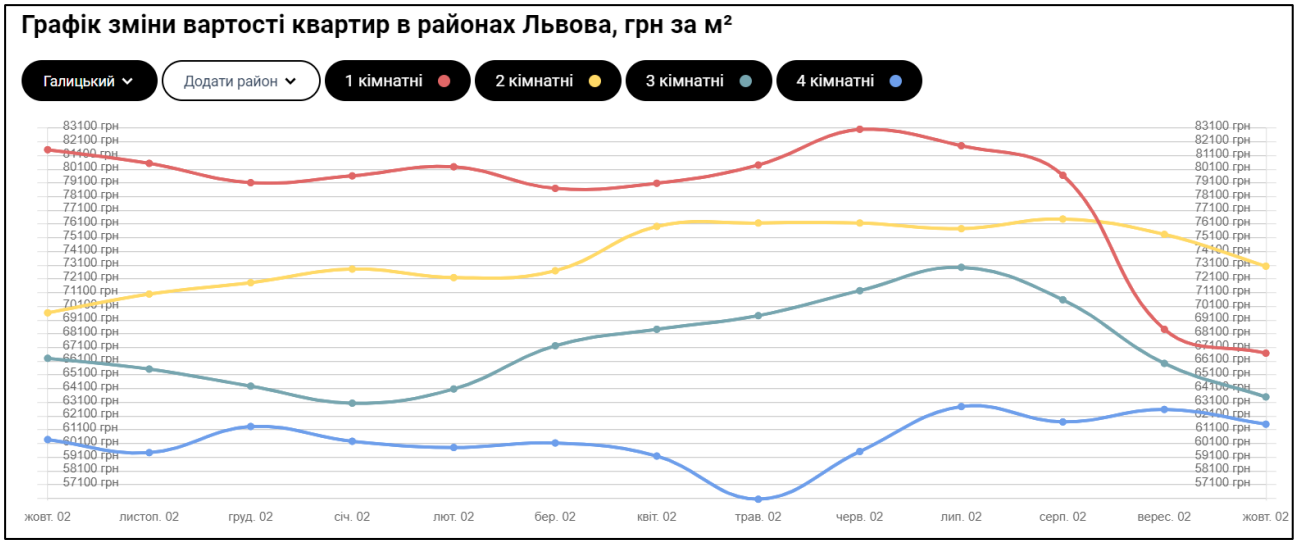

Рисунок 1.7 – Графік зміни вартості квартир в Галицькому районі Львова

У Шевченківському районі вартість новобудов становить приблизно **58 030 грн** за квадратний метр (див. рис. 1.8).

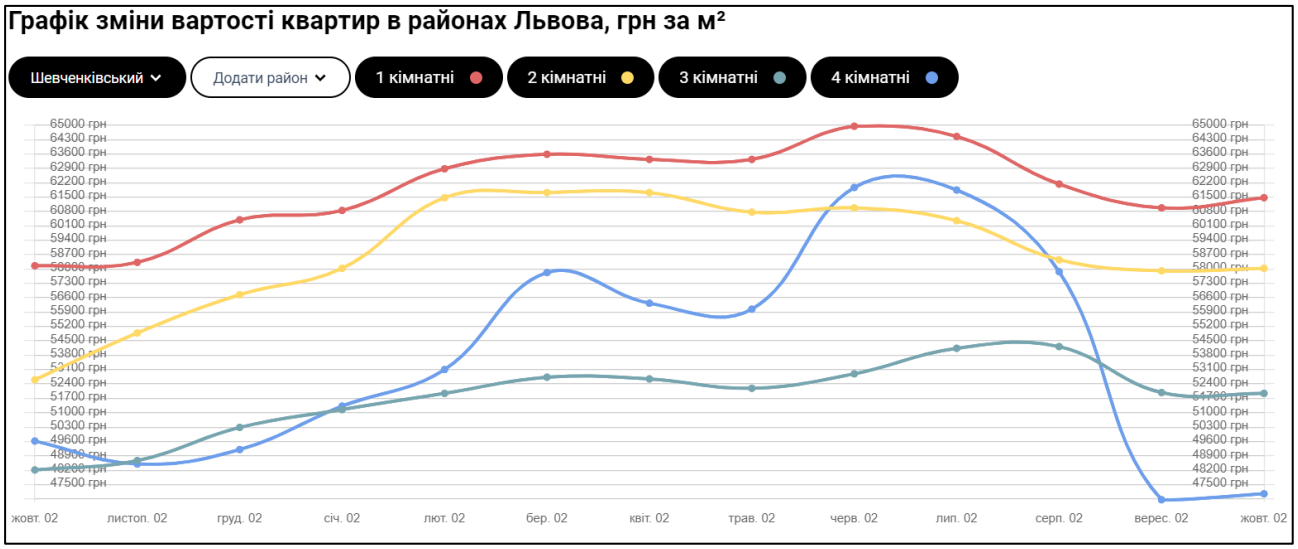

Рисунок 1.8 – Графік зміни вартості квартир в Шевченківському районі Львова

Для елітних новобудов ціни можуть перевищувати **88 274 грн** за квадратний метр, що характерно для будинків вищого класу. У той час, квартири економ-класу пропонуються за цінами від **37 261 грн** за квадратний метр. Такі ціни обумовлені рядом факторів, включаючи інфраструктуру районів, доступність до центрів міста, розвиток соціальних і транспортних мереж, а також попит на комфортне житло у Львові [7-8].

Аналіз ринку нерухомості у Львові показує, що попит на житло залишається високим завдяки культурно-історичним, економічним та соціальним факторам.

1.4 Аналіз інформаційних систем та сучасних методів прогнозування

Прогнозування цін на нерухомість є надзвичайно важливим аспектом сучасної економіки, оскільки воно впливає на рішення як окремих осіб, так і великих інституцій. На ринку нерухомості ціни визначаються не лише характеристиками самих об'єктів, такими як площа, кількість кімнат або стан нерухомості, але й цілою низкою зовнішніх факторів, зокрема економічними умовами, соціальною ситуацією, змінами в законодавстві та інфраструктурними проектами. Тому точне прогнозування цін може допомогти учасникам ринку

ухвалювати більш обґрунтовані рішення, уникати фінансових втрат та максимізувати прибутки [9].

Сучасні методи прогнозування цін на нерухомість охоплюють різноманітні підходи, включаючи традиційні статистичні методи, такі як множинний регресійний аналіз, а також новітні технології машинного навчання, зокрема штучні нейронні мережі. Використання алгоритмів машинного навчання дозволяє враховувати складні залежності та обробляти великі обсяги даних, що значно підвищує точність прогнозів. Наприклад, з розвитком технологій автоматизовані моделі оцінки вартості стають дедалі більш доступними, що робить їх корисними інструментами для аналізу ринку.

У статті "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times" автори досліджують ефективність різних алгоритмів машинного навчання для прогнозування цін на житло, зокрема в умовах впливу пандемії COVID-19 на ринок нерухомості. Основна увага приділяється порівнянню ансамблевих алгоритмів, таких як Gradient Boosting Regressor (GBR), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), а також традиційним методам, таким як лінійна регресія [9-10].

Gradient Boosting Regressor працює на основі методу бустингу, який будує моделі послідовно, кожна з яких намагається виправити помилки попередньої. Цей підхід дозволяє ефективно адаптуватися до складних даних, виявляючи важливі закономірності, що робить його потужним інструментом для прогнозування [11].

Extreme Gradient Boosting, або XGBoost, є вдосконаленою версією GBR, яка використовує оптимізації для підвищення швидкості навчання та зменшення ризику перенавчання. Цей алгоритм здатний обробляти великі обсяги даних і демонструє високу точність у прогнозах, що робить його популярним у багатьох аналітичних задачах.

Light Gradient Boosting Machine (LGBM) також належить до класу бустингових алгоритмів, але оптимізований для швидшої обробки та меншого

споживання пам'яті. LGBM реалізує нові підходи до обробки даних, що дозволяє ефективніше працювати з великими наборами даних, що є звичайною практикою на ринку нерухомості [11].

У статті також обговорюється використання традиційної лінійної регресії як базового методу для порівняння. Хоча цей метод є простим та зрозумілим, його обмеження стають очевидними в контексті складних ринкових даних, де нелінійні взаємозв'язки можуть суттєво вплинути на точність прогнозів (див. табл. 1.4).

Таблиця 1.4. – Алгоритми машинного навчання використані у статті

Id	**Назва**	**Модель**	**Бібліотека**
1	*lr*	Linear Regression	sklearn.linear_model.LinearRegression
2	*rf*	Random Forest Regressor	sklearn.ensemble. RandomForestRegressor
3	*et*	Extra Trees Regressor	sklearn.ensemble.ExtraTreesRegressor
4	*gbr*	Gradient Boosting Regressor	sklearn.ensemble. GradientBoostingRegressor
5	*xgbm*	Extreme Gradient Boosting	xgboost.XGBRegressor
6	*lgbm*	Light Gradient Boosting Machine	lightgbm.LGBMRegressor

Загалом, результати дослідження свідчать про те, що машинні алгоритми навчання, зокрема XGBoost та LGBM, демонструють значні переваги в точності прогнозування цін на житло в порівнянні з традиційними підходами (див. табл. 1.5).

Таблиця 1.5 – Результати продуктивності навчених алгоритмів (оцінка R^2).

Модель	**Назва**	**Крос-валідація в навчальному наборі (СД)**	**R^2**		
			Навчальна вибірка	**Тестова вибірка**	**Перенавчання (%)**
Linear Regression	*lr*	0.8048 (0.0060)	0.8056	0.8052	-
Random Forest Regressor	*rf*	0.9036 (0.0049)	0.9970	0.9135	+9.1
….....					

Кінець таблиці 1.5

Модель	**Назва**	**Крос-валідація в навчальному наборі (СД)**	**R^2**		
			Навчальна вибірка	**Тестова вибірка**	**Перенавчання (%)**
Extra-Trees Regressor	*et*	0.9101 (0.0040)	0.9997	0.9178	+8.9
Gradient Boosting Regressor	*gbr*	**0.9125** (0.0034)	0.9952	**0.9192**	**+8.3**
Extreme Gradient Boosting	*xgbm*	0.9094 (0.0041)	0.9900	0.9178	+7.9
Light Gradient Boosting Machine	*lgbm*	0.9076 (0.0044)	0.9902	0.9140	+8.3

У статті "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times" представлено детальний процес навчання, оптимізації, оцінювання та вибору моделей, що ілюструється на графічному зображенні (див. рис. 1.9).

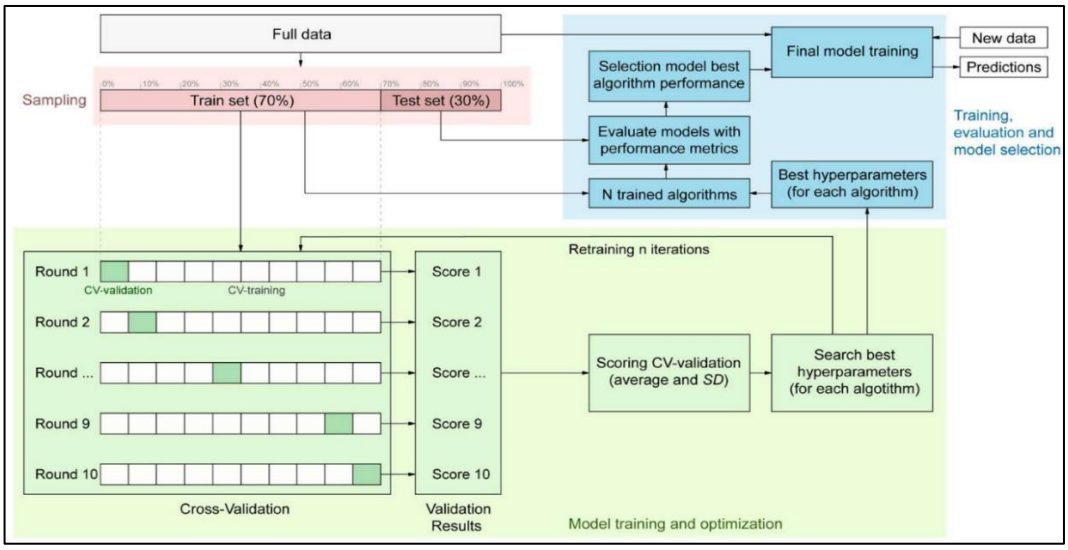

Рисунок 1.9 – Робочі процеси на етапах навчання, оптимізації, оцінювання та вибору моделі [10]

На початку етапу обробки даних здійснюється **вибірка** з повного набору даних, де дані діляться на навчальну (70%) та тестову (30%) вибірки. Цей етап є критично важливим, оскільки правильне розподілення даних забезпечує надійність моделей під час оцінювання їхньої продуктивності [10].

Наступним етапом є **крос-валідація**, де навчальна вибірка поділяється на кілька раундів. Наприклад, на схемі зазначено 10 раундів крос-валідації, в яких дані розподіляються на підвибірки для навчання та валідації. У кожному раунді модель тренується на одній частині даних, тоді як інша частина використовується для перевірки її точності. Цей процес дозволяє отримати різні оцінки (Scores) ефективності моделі для кожного з раундів, що забезпечує більш об’єктивний аналіз її продуктивності [10].

Після проведення крос-валідації результати оцінок аналізуються і на їх основі вибираються найкращі гіперпараметри для кожного алгоритму. Це є важливим етапом, оскільки оптимізація гіперпараметрів може значно підвищити точність прогностичних моделей.

Наступний крок передбачає **навчання фінальної моделі** з урахуванням отриманих найкращих гіперпараметрів. Після цього система готова до застосування на нових даних, де проводиться прогнозування цін на нерухомість.

Завдяки такій структурі процесу навчання, оптимізації та вибору моделей, автори статті демонструють ефективний підхід до прогнозування цін на житло, який може бути адаптований для використання в реальних умовах ринку нерухомості. Цей підхід також підкреслює важливість ретельного підбору алгоритмів і параметрів для досягнення високої точності у прогнозах, що є критично важливим для прийняття обґрунтованих рішень у сфері нерухомості.

У статті "House Price Prediction Using Machine Learning" автор Ченксі Лі аналізує застосування різних алгоритмів машинного навчання для прогнозування цін на житло в окрузі Кінг, штат Вашингтон [12]. Дослідження зосереджується на кількох ключових алгоритмах, які використовуються для оцінки вартості

нерухомості, включаючи лінійну регресію, випадковий ліс, нейронні мережі та XGBoost.

Лінійна регресія була використана як базовий метод для порівняння з більш складними моделями. Цей метод дозволяє встановити лінійний зв'язок між залежною змінною (ціною) та незалежними змінними (такі як площа, кількість кімнат тощо). Однак автор зазначає, що лінійна регресія може бути недостатньою для адекватного моделювання складних залежностей у даних, що є характерним для ринку нерухомості, де присутні нелінійні зв'язки.

Випадковий ліс (Random Forest) є ансамблевим методом, що об'єднує результати кількох дерев рішень. Цей алгоритм має переваги в обробці великих обсягів даних і демонструє високу стійкість до перенавчання завдяки своїй здатності враховувати різноманітні фактори. У статті було показано, що випадковий ліс забезпечує більш точні прогнози, ніж лінійна регресія, завдяки своїй здатності адаптуватися до складних, нелінійних зв'язків у даних [13-14].

XGBoost, або екстремальний градієнтний бустинг, є ще одним потужним інструментом, що використовується в дослідженні. Цей алгоритм оптимізує процес навчання, дозволяючи швидко обробляти великі набори даних і знижувати ризик перенавчання. XGBoost продемонстрував високу точність в прогнозах, завдяки своїм адаптивним властивостям і можливостям регуляризації, що робить його ідеальним для аналізу ринку нерухомості (див. рис. 1.10).

Нейронні мережі були також використані в дослідженні як один з найсучасніших підходів до прогнозування. Вони здатні навчатися на великій кількості даних та виявляти складні патерни, які можуть бути непомітні для традиційних методів. Автор зазначає, що нейронні мережі продемонстрували значне покращення точності прогнозів у порівнянні з класичними методами [15] (див. табл. 1.6).

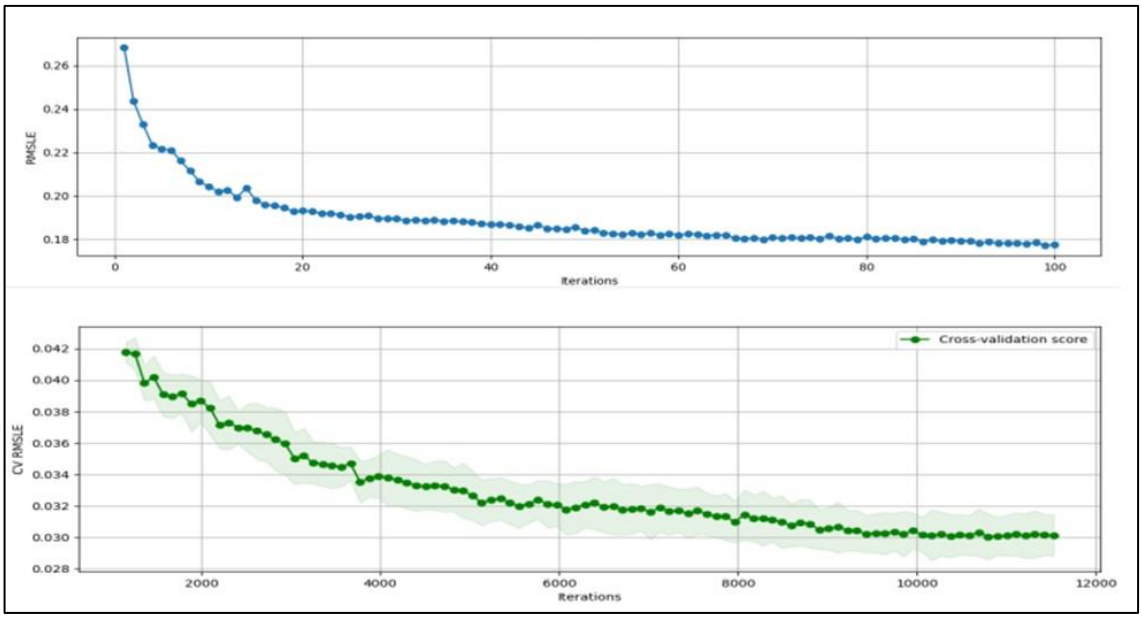

Рисунок 1.10 – Результати дослідження для Random Forest (перша крива) та XGBoost (друга крива)

Таблиця 1.6 – Порівняння точності моделей [15]

Model type	R-squared	RMSE	MSE
Linear regression	0.706	210649.7710	44373326009.8143
Random forest	0.878	136170.2574	18542338996.4575
Neural network	0.846	143075.1230	22825808222.9256
XGBoost	0.888	130281.3637	16973233719.7849

Загалом, стаття Ченксі Лі підкреслює, що машинне навчання та використання алгоритмів, таких як випадковий ліс, нейронні мережі та XGBoost, дозволяє досягти вищої точності і надійності прогнозів, що є критично важливим для ухвалення обґрунтованих рішень на ринку нерухомості.

Аналіз інформаційних систем та сучасних методів прогнозування цін на нерухомість вказує на значні зміни у підходах до оцінки вартості житлових об'єктів, завдяки інтеграції новітніх алгоритмів машинного навчання. Від традиційних статистичних методів, таких як лінійна регресія, до сучасних ансамблевих алгоритмів, таких як Gradient Boosting, XGBoost та Light Gradient Boosting Machine, спостерігається значне підвищення точності прогнозів.

Сучасні підходи дозволяють враховувати складні нелінійні взаємозв'язки в даних, що традиційні моделі не можуть достатньо точно відобразити. Дослідження демонструють, що алгоритми машинного навчання, зокрема XGBoost і LGBM, переважають традиційні методи в контексті точності прогнозування, що має критичне значення для прийняття рішень на ринку нерухомості.

Ці результати свідчать про те, що впровадження сучасних технологій в інформаційні системи для прогнозування цін на нерухомість може суттєво поліпшити якість оцінок, зменшити ризики для інвесторів та забезпечити більш обґрунтовані рішення для учасників ринку. Це також відкриває нові можливості для автоматизації процесів оцінки нерухомості, що є важливим кроком для подальшого розвитку ринку. Таким чином, об'єднання традиційних та сучасних підходів створює більш ефективні та надійні механізми для прогнозування цін на нерухомість, що може стати основою для подальших досліджень і практичних застосувань у цій галузі.

1.5 Постановка задачі

Загальна проблематика питання полягає в необхідності розробки інформаційної системи для прогнозування цін на ринку нерухомості, яка б дозволяла автоматизувати процес оцінки вартості об'єктів на основі різноманітних критеріїв. В умовах високої динамічності цього ринку важливо мати інструмент, здатний обробляти великі обсяги даних та враховувати численні фактори, що впливають на ціноутворення, такі як розташування, стан нерухомості, економічні умови та інші параметри.

Актуальність теми полягає в необхідності точного прогнозування цін на ринку нерухомості для підтримки прийняття обґрунтованих рішень в умовах високої динамічності цього ринку. Використання сучасних методів машинного навчання дозволяє значно підвищити точність прогнозів і автоматизувати процес оцінки вартості нерухомості. Розробка програмного забезпечення для цієї задачі

дасть можливість за допомогою інтерактивного інтерфейсу здійснювати швидкий і точний розрахунок вартості об'єктів на основі ключових факторів, що впливають на ціноутворення. Це, в свою чергу, дозволить знизити ризики при прийнятті рішень на ринку нерухомості та підвищити ефективність інвестиційних стратегій.

Об’єктом дослідження є процес прогнозування цін на ринку нерухомості за допомогою інформаційної системи, що аналізує різноманітні фактори, які впливають на вартість нерухомості.

Предметом дослідження є інформаційна система, що здійснює прогнозування цін на ринку нерухомості. Вона включає в себе методи та алгоритми аналізу даних, які використовуються для обробки інформації про об'єкти нерухомості та фактори, що впливають на формування цін.

Метою роботи є прогнозування цін на ринку нерухомості з використанням інформаційної системи, яка здійснює аналіз різноманітних факторів, що впливають на вартість об'єктів.

Завдання для досягнення поставленої мети:

– аналіз сучасних інформаційних систем та методів прогнозування;

– аналіз загальної структури інформаційної системи прогнозування цін на ринку нерухомості;

– порівняльний аналіз отриманих результатів для визначення найбільш ефективних методів прогнозування;

– реалізація інформаційної системи прогнозування цін на ринку нерухомості.

Висновки до розділу 1

У першому розділі кваліфікаційної роботи проведено комплексний аналіз ринку нерухомості та сучасних методів прогнозування цін. Дослідження розпочалося з характеристики ринку нерухомості, що включає ключові елементи, такі як попит і пропозиція, регіональні особливості та вплив соціально-

економічних факторів. Визначено, що ринок нерухомості є динамічною системою, чутливою до змін в економіці, законодавстві та суспільних умовах.

Аналіз факторів, що впливають на ціноутворення, підкреслив, що для отримання точних прогнозів необхідно враховувати не лише характеристики об'єктів, але й широкий спектр зовнішніх впливів. Сучасні інформаційні системи, що використовують методи машинного навчання, демонструють значну перевагу в точності порівняно з традиційними статистичними методами.

У розділі також було розглянуто основні алгоритми машинного навчання, такі як Gradient Boosting, XGBoost та LGBM, які довели свою ефективність у прогнозуванні цін на нерухомість. Проаналізовані статті продемонстрували, що ці алгоритми здатні адаптуватися до складних, нелінійних даних, що є характерними для ринку нерухомості.

Таким чином, можна зробити висновок про необхідність впровадження сучасних технологій в інформаційні системи для прогнозування цін на нерухомість. Це не лише підвищує точність оцінок, але й відкриває нові можливості для автоматизації процесів оцінки, що є критично важливим для учасників ринку.

2 СТРУКТУРА ІНФОРМАЦІЙНОЇ СИСТЕМИ ТА МЕТОДИ ЇЇ РЕАЛІЗАЦІЇ

2.1 Структура інформаційної системи

Структура інформаційної системи прогнозування цін на ринку нерухомості побудована на трьох основних етапах, кожен з яких є важливим для досягнення точності прогнозів (див. рис. 2.1).

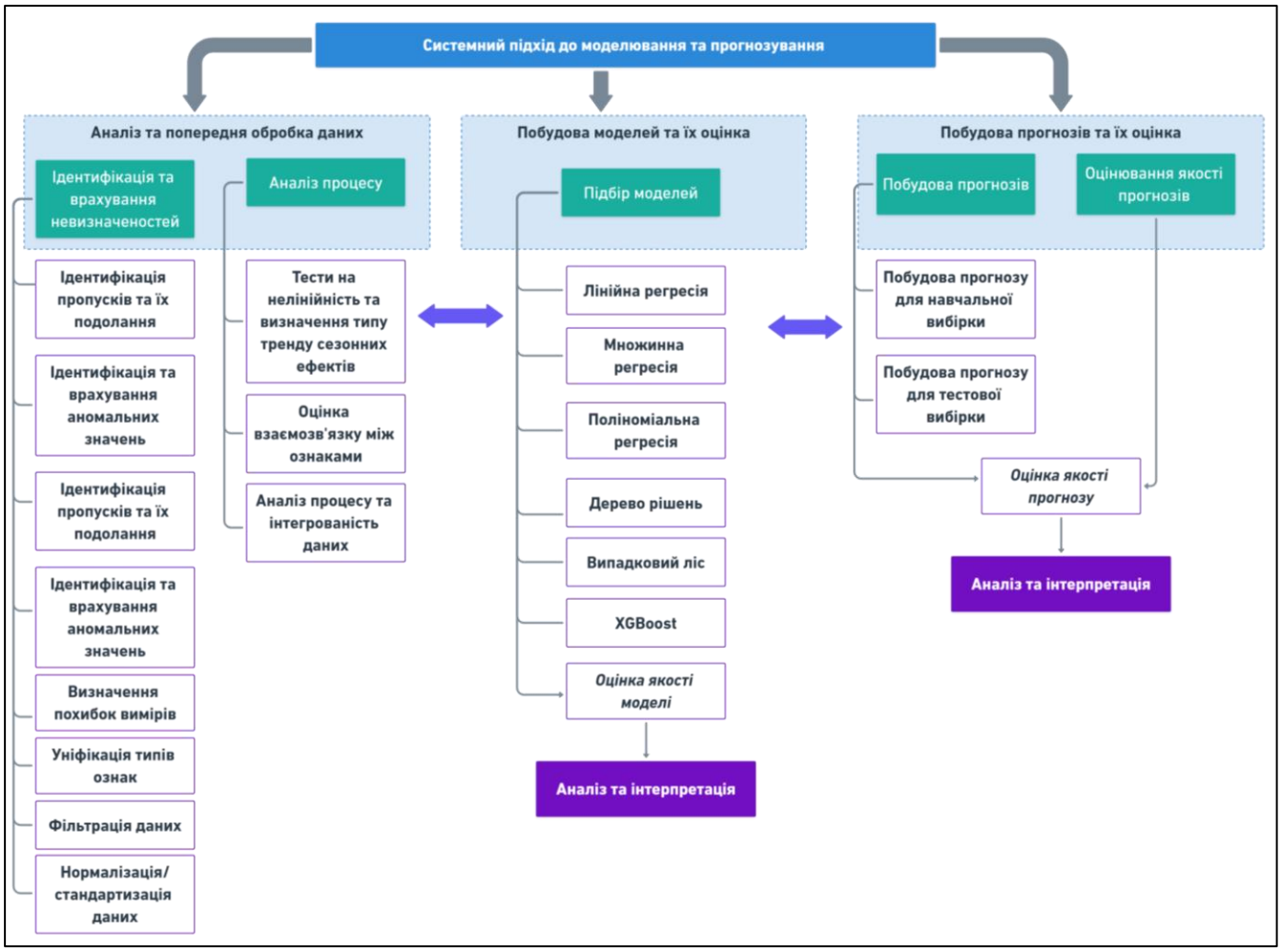

Рисунок 2.1 – Структура інформаційної системи прогнозування

Перший етап інформаційної системи прогнозування цін на ринку нерухомості зосереджений на попередній обробці даних. Він є критичним для забезпечення якості та точності результатів моделювання, оскільки погано підготовлені або

невірно оброблені дані можуть значно вплинути на ефективність роботи системи в цілому.

Процес попередньої обробки даних включає кілька важливих кроків, кожен з яких має на меті привести набір даних до оптимального виду для подальшого використання в алгоритмах машинного навчання. Спочатку відбувається ідентифікація та врахування невизначеностей у даних. Це означає, що всі потенційно неточні або неповні дані мають бути знайдені і позначені. Це можуть бути, наприклад, пропущені значення або такі аномалії, як значення, які виходять за межі нормального діапазону для кожної з характеристик [16].

Далі здійснюється коригування аномальних значень. Виявлені аномалії можуть бути виключені зі складу даних, або ж для них застосовуються спеціальні методи заміни, такі як середнє значення, медіана, або ж інші методи, залежно від природи пропусків. Цей крок дозволяє мінімізувати вплив помилок на модель і підвищити її точність [17].

Окрім того, для кожної змінної, що входить до аналізу, може бути здійснено нормалізацію або стандартизацію даних. Нормалізація передбачає приведення значень змінних до певного діапазону, наприклад, між 0 і 1, що особливо важливо, якщо використовуються алгоритми, чутливі до масштабування даних, такі як нейронні мережі або деякі типи регресії. Стандартизація, в свою чергу, допомагає привести розподіл даних до стандартного виду з середнім значенням, що дорівнює нулю, і стандартним відхиленням, яке дорівнює одиниці. Цей крок є важливим для алгоритмів, які чутливі до різних масштабів змінних, зокрема для лінійної регресії [18].

Важливим етапом попередньої обробки є також уніфікація типів ознак. У цьому випадку всі категоріальні ознаки повинні бути переведені в числові значення, що дозволяє їх використання в математичних моделях. Для цього застосовуються методи, такі як кодування категоріальних змінних через індекси.

Завершальним кроком на цьому етапі є фільтрація даних, яка дозволяє усунути записи, що не відповідають встановленим критеріям або які містять некоректні дані. Наприклад, записи, де відсутні ключові параметри (ціна або площа) можуть бути виключені з аналізу, оскільки їх наявність є критично важливою для побудови моделі [19].

Таким чином, перший етап попередньої обробки даних забезпечує надійну основу для наступних етапів моделювання. Він допомагає уникнути проблем, пов'язаних з некоректними або неповними даними, та забезпечує більш високу точність прогнозів.

Другий етап інформаційної системи прогнозування цін на ринку нерухомості спрямований на побудову та оцінку моделей. Це один із найважливіших етапів, оскільки правильний вибір моделі та її налаштування визначають здатність системи надавати точні прогнози. Після того, як дані були попередньо оброблені, система може переходити до етапу моделювання, на якому використовуються різноманітні алгоритми машинного навчання для побудови моделей, здатних прогнозувати ціни на основі вхідних характеристик.

На цьому етапі були побудовані кілька різних моделей для прогнозування цін на ринку нерухомості, зокрема, лінійна регресія, множинна регресія, поліноміальна регресія, дерева рішень, випадковий ліс і XGBoost. Кожна з цих моделей має свої особливості та переваги в залежності від складності та виду даних. Моделі, побудовані за допомогою регресійних методів, дозволяють прогнозувати значення залежної змінної (ціни) на основі вхідних параметрів, таких як площа, стан ремонту, кількість кімнат та інші фактори [20].

Лінійна регресія є базовим методом, який використовує лінійну залежність між незалежними змінними та залежною. Цей метод є швидким і простим у реалізації, але його недоліком є обмежена здатність моделювати нелінійні зв'язки між змінними.

Множинна регресія розширює лінійну регресію за рахунок використання кількох незалежних змінних для прогнозування. Це дозволяє моделювати більш складні взаємозв'язки між різними характеристиками об'єкта нерухомості [21].

Поліноміальна регресія є варіантом множинної регресії, який дозволяє моделювати нелінійні залежності шляхом додавання поліноміальних членів до моделі. Це дає змогу досягти більш високої точності при прогнозуванні цін у випадках, коли зв'язки між характеристиками об'єкта не є лінійними.

Дерева рішень є інтерпретованим методом, який використовує розбиття даних на групи на основі умов, що дозволяє прогнозувати ціни в залежності від конкретних значень характеристик. Цей метод дуже добре працює з категоріальними даними і надає ясну інтерпретацію результатів [22-23].

Випадковий ліс є ансамблевим методом, який використовує кілька дерев рішень для побудови прогнозу, що дозволяє покращити точність і стабільність моделі порівняно з окремим деревом [24].

XGBoost (Extreme Gradient Boosting) є ще одним потужним ансамблевим методом, що базується на методі градієнтного бустингу і використовується для оптимізації прогностичних моделей. Цей метод є одним з найефективніших при роботі з великими наборами даних, оскільки він виявляє складніші патерни, ніж інші методи [25].

Після побудови моделей необхідно здійснити їх оцінку, щоб визначити, яка з них є найкращою для даного набору даних. Для оцінки якості моделей прогнозування використовуються різноманітні статистичні метрики, які дозволяють оцінити точність та ефективність моделі. Однією з таких метрик є критерій Акаїке (AIC), який допомагає оцінити компроміс між точністю моделі та її складністю. Він дозволяє вибрати найбільш оптимальну модель, мінімізуючи надмірну складність, що може призвести до перенавчання. Чим менше значення AIC, тим краща модель для даного набору даних.

Іншою важливою метрикою є критерій Дарбіна-Уотсона (DW), який використовується для перевірки наявності автокореляції у залишках моделі. Якщо значення DW близьке до 2, це вказує на відсутність автокореляції, що є ознакою хорошої моделі. Якщо значення відхиляється від 2 в бік 0 або 4, це може свідчити про проблеми з автокореляцією залишків, що потребує коригування моделі.

Статистика Фішера (F) є ще однією важливою метрикою, яка використовується для оцінки значущості регресійної моделі в цілому. Вона дозволяє перевірити гіпотезу про те, чи є в моделі достатньо статистичної сили для пояснення варіацій залежної змінної. Високе значення F вказує на значущість моделі, а низьке значення може свідчити про її недостатню ефективність [26].

Крім того, оцінка моделі може включати аналіз R-квадрат – показника, що вказує на частку варіації залежної змінної, яку можна пояснити незалежними змінними в моделі. Високе значення R-квадрат свідчить про високу здатність моделі до пояснення варіацій у даних.

Використовуючи різні метрики оцінки, система може вибрати найбільш ефективну модель для прогнозування цін на ринку нерухомості. Крім того, після побудови та оцінки моделей, система може бути готова до виконання прогностичних завдань, таких як прогнозування цін для нових об'єктів нерухомості на основі наданих користувачем даних.

Таким чином, другий етап є важливим, оскільки він безпосередньо відповідає за створення моделей, здатних адекватно прогнозувати ціни, а також за їх належну оцінку для подальшого використання в реальному застосуванні.

Третій етап інформаційної системи прогнозування цін на ринку нерухомості спрямований на побудову та оцінку прогнозів, що є кінцевим результатом роботи моделі на практиці. Цей етап важливий, оскільки визначає, наскільки ефективно система може бути використана для реальних прогнозувань на ринку нерухомості, допомагаючи користувачам приймати обґрунтовані рішення на основі отриманих даних.

Після того, як дані були попередньо оброблені та побудовані відповідні прогностичні моделі (множинної регресії, поліноміальної регресії, дерев рішень, випадковий ліс, XGBoost), наступним кроком є застосування побудованих моделей для прогнозування цін. Це відбувається шляхом введення нових даних (наприклад, характеристик нерухомості, таких як площа, кількість кімнат, стан ремонту та інші параметри) в систему, яка потім використовує модель для генерації прогнозу ціни для конкретного об'єкта.

Основною метою третього етапу є формування прогнозу, який можна використовувати для оцінки ринкової вартості нерухомості, аналізу змін в умовах ринку або порівняння потенційних інвестиційних можливостей. В системі цей процес автоматизовано, що дозволяє користувачеві легко вводити необхідні дані, а система виводить прогнозовану ціну.

Для того, щоб система була ефективною та забезпечувала точні прогнози, важливо, щоб моделі мали високу точність прогнозування. Це досягається через постійну перевірку точності моделі, а також через коригування параметрів моделі в разі необхідності. Для перевірки точності прогнозів використовуються така метрика, як середня квадратична помилка (MSE), яка дає уявлення про середнє відхилення прогнозованих значень від реальних.

Водночас на третьому етапі оцінюється якість прогнозів, що є важливою частиною для забезпечення ефективності системи в реальних умовах. За допомогою аналізу результатів тестування моделі та оцінки її показників, можна зробити висновки про її готовність до застосування на практиці. У разі виявлення проблем або низької точності, можуть бути внесені коригування до моделі або здійснено її доопрацювання, що дозволить покращити точність прогнозів.

Також на цьому етапі здійснюється аналіз та інтерпретація результатів прогнозу. Це означає, що система не тільки генерує прогнозовані ціни, але й надає користувачеві змогу зрозуміти, які фактори найбільше впливають на вартість

нерухомості, що дозволяє робити висновки про важливість певних характеристик об'єкта для формування його ринкової ціни.

Загалом, третій етап є вирішальним для застосування моделі в реальних умовах, оскільки дозволяє отримувати конкретні прогнози, які можуть використовуватись для практичних цілей у сфері нерухомості, таких як оцінка вартості об'єктів або аналіз ринкових тенденцій.

Процес взаємодії користувача з системою прогнозування цін на ринку нерухомості розпочинається з введення необхідних даних у інтерфейс системи. Користувач надає параметри, такі як площа, кількість кімнат, стан ремонту та інші характеристики об'єкта нерухомості. Після цього система автоматично обробляє ці дані, застосовує модель множинної регресії та здійснює прогнозування ціни на основі введених параметрів. Результат прогнозування виводиться на екран, що дозволяє користувачу отримати орієнтовну ціну об'єкта на ринку нерухомості, що є зручним і швидким інструментом для оцінки вартості майна (див. рис. 2.2).

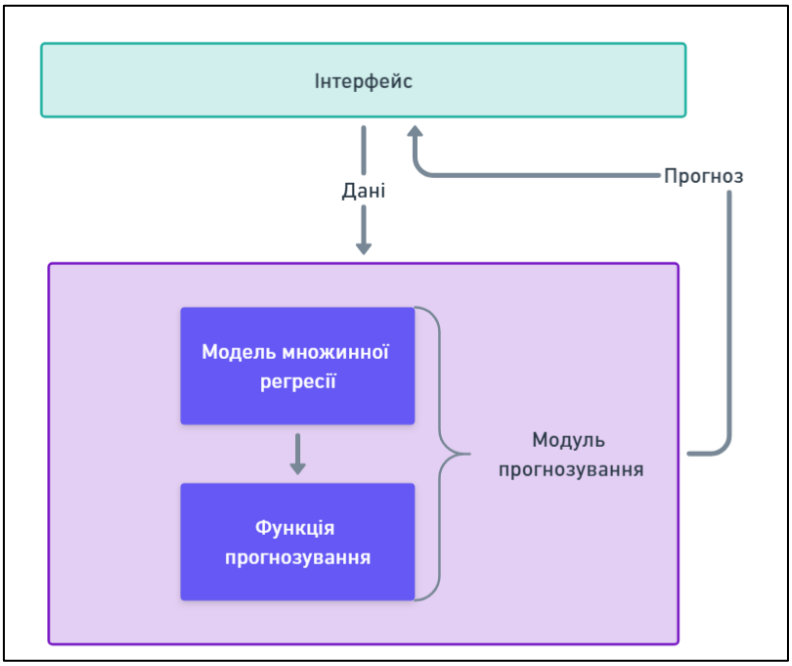

Рисунок 2.2 – Процес взаємодії користувача з системою прогнозування цін на ринку нерухомості

2.2 Технології розробки системи

Мова програмування R була обрана для розробки інформаційної системи прогнозування цін на ринку нерухомості завдяки її потужним можливостям для статистичного аналізу та обробки даних. R є однією з найбільш популярних мов для виконання складних обчислень і аналізу великих обсягів даних, що робить її ідеальним інструментом для побудови моделей прогнозування. Вона надає велику кількість бібліотек для реалізації алгоритмів машинного навчання, таких як лінійна регресія, випадковий ліс, XGBoost, що дозволяє будувати точні моделі для прогнозування цін на основі різних характеристик нерухомості. Крім того, R має потужні засоби для візуалізації результатів, такі як бібліотека ggplot2, що допомагає створювати графіки для аналізу та інтерпретації даних. Важливою перевагою є інтеграція з іншими технологіями, що дозволяє системі легко адаптуватися до великих даних та різних обчислювальних середовищ. Використання RStudio як середовища розробки забезпечує зручність для програмування, відлагодження та тестування моделей, а також дозволяє створювати інтерактивні додатки для кінцевих користувачів за допомогою бібліотеки Shiny. У результаті вибір R гарантує високу точність, гнучкість і зручність у розробці та експлуатації системи прогнозування цін на ринку нерухомості.

Розробка інформаційної системи прогнозування цін на ринку нерухомості здійснюватиметься в середовищі RStudio, що є потужним інтегрованим середовищем розробки для мови програмування R. Це середовище забезпечує всі необхідні інструменти для написання, тестування та налагодження коду, що дозволяє ефективно працювати з великими обсягами даних та застосовувати складні моделі машинного навчання. RStudio підтримує інтерактивну роботу з даними, що полегшує процес аналізу та візуалізації результатів, а також має зручні функції для налаштування та оптимізації моделей. Використання RStudio дозволяє швидко інтегрувати різні бібліотеки для аналізу, побудови моделей і створення візуалізацій, що значно спрощує розробку і дає можливість зосередитися на

вирішенні основних завдань системи. Середовище також має вбудовані інструменти для роботи з графіками та звітами, що допомагає візуалізувати дані та результати моделювання. Завдяки своїй зручності, RStudio дозволяє оптимізувати весь процес розробки, забезпечуючи високу продуктивність і ефективність на всіх етапах створення системи.

Для реалізації інтерфейсу взаємодії користувача з системою прогнозування цін на ринку нерухомості буде використано методи та бібліотеки, які забезпечують інтерактивність і зручність використання. Однією з основних бібліотек для створення інтерфейсу, є **Shiny**, що є потужним інструментом для створення веб-додатків без необхідності глибоких знань веб-програмування. Ця бібліотека дозволяє швидко розробити динамічний інтерфейс, через який користувач зможе вводити параметри, такі як площа, кількість кімнат, стан ремонту та інші характеристики об'єкта нерухомості. Після введення даних користувачем, система обробляє їх за допомогою моделей машинного навчання, таких як множинна регресія, та виводить прогнозовану ціну на екран. Такий підхід дозволяє створити інтуїтивно зрозумілий інтерфейс для кінцевого користувача, що робить систему зручною та доступною для широкого кола людей, включаючи тих, хто не має спеціальних знань у галузі даних або програмування.

Висновки до розділу 2

У другому розділі кваліфікаційної роботи було детально описано структуру інформаційної системи прогнозування цін на ринку нерухомості, а також методи її розробки. В рамках системи були визначені ключові етапи, такі як попередня обробка даних, побудова моделей та їх оцінка, а також реалізація інтерактивного інтерфейсу для взаємодії з користувачем. Основною метою системи є точне прогнозування цін на основі різних факторів, що впливають на ринок нерухомості.

Для розробки було обрано середовище RStudio та мову програмування R, що дозволяє ефективно працювати з даними, застосовувати різноманітні методи

машинного навчання та створювати графічні інтерфейси для користувачів. Використані бібліотеки, такі як ggplot2, caret, xgboost та інші, дозволять реалізувати необхідні функціональності для моделювання, візуалізації та оцінки прогнозів.

Завдяки інтеграції цих технологій буде створена система, здатна надавати точні прогнози цін на ринку нерухомості в зручному та доступному форматі для кінцевих користувачів.

3 АНАЛІЗ ТА ПОПЕРЕДНЯ ОБРОБКА ДАНИХ

3.1 Алгоритм аналізу та очищення даних

Для реалізації інформаційної системи прогнозування цін на ринку нерухомості критично важливим етапом є попередня обробка даних. Цей етап забезпечує підготовку даних для ефективного навчання моделей, що значно впливає на точність прогнозів.

Процедуру аналізу та очищення даних виконно на основі алгоритму, який представлено у вигляді блок-схеми на рисунку 3.1.

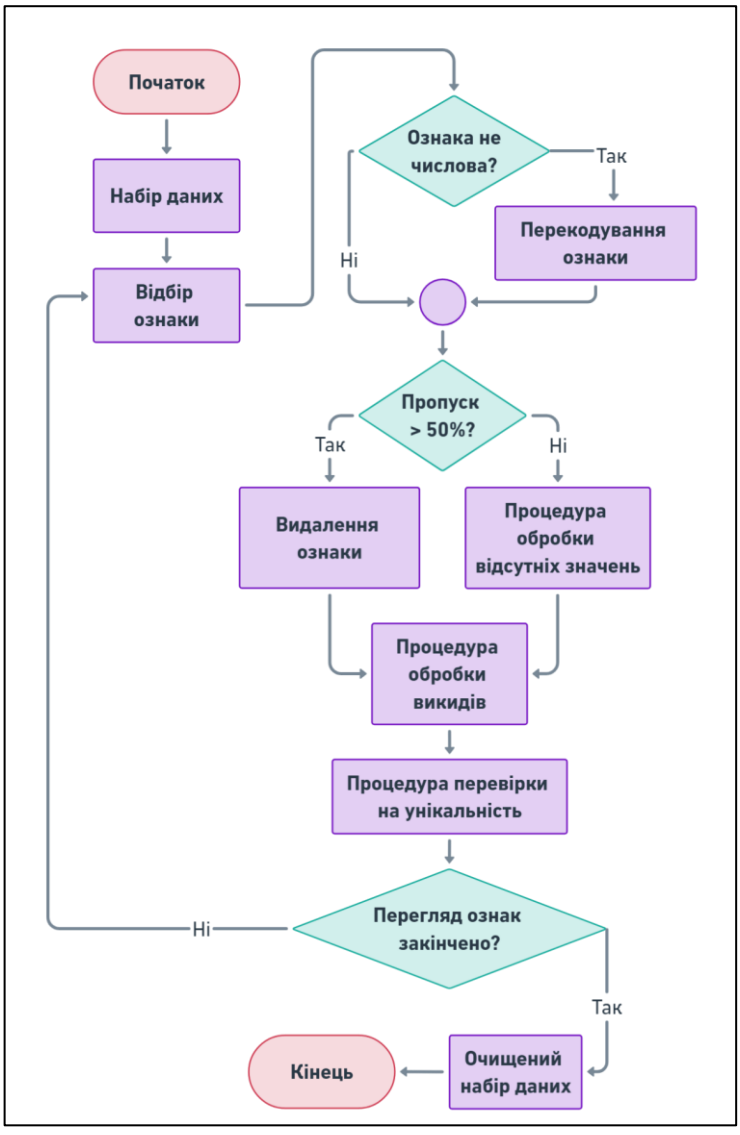

Рисунок 3.1 – Блок-схема алгоритму очищення даних

Алгоритм очищення даних починається з етапу завантаження даних, де дані імплементуються для подальшого аналізу. Наступним кроком є вибір ознак, у процесі якого визначаються змінні, що будуть використовуватися в подальшому моделюванні [27-29].

Після цього проводиться оцінка типу обраних ознак. Якщо ознака є числовою, перехід до наступного етапу, де перевіряється наявність пропусків. Якщо пропусків більше 50%, ознака видаляється з набору даних. У випадку, коли пропусків менше 50%, проводиться процедура обробки відсутніх значень.

Далі алгоритм включає обробку викидів, що передбачає виявлення аномальних значень у даних, які можуть впливати на точність моделі. Після цього виконується процедура перевірки на унікальність, що включає виявлення дублікатів у наборі даних.

Останнім етапом є оцінка завершення перевірки ознак. Якщо всі обрані ознаки були перевірені, алгоритм завершує свою роботу, а очищений набір даних зберігається для подальшого використання у прогнозуванні [27-29]. Таким чином, даний алгоритм забезпечує надійність і точність даних, що є критично важливими для створення ефективної інформаційної системи.

3.2 Попередня обробка даних

Для виконання практичної частини дипломної роботи було використано, як приклад, імітаційний набор даних flats.csv, що містить дані щодо вартості та характеристик квартир.

У файлі перераховано ціни квартир, тип, метраж, стан, локація та кількість кімнат (див. рис. 3.2).

	A	B	C	D	E	F	G
1	rooms	location	condition	m2	type	price	
2	2	suburbs	repaired	50	used	35000	
3	1	center	repaired	37	used	35000	
4	3	suburbs	repaired	67	used	65000	
5	NA	suburbs	repaired	21	used	15000	
6	1	suburbs	repaired	82	NA	60000	
7	3	center	repaired	82	used	85000	
8	2	center	repaired	45	used	48000	
9	3	center	repaired	82	used	85000	
10	1	suburbs	unrepaire	41	new	30000	

Рисунок 3.2 – Фрагмент набору даних flats.csv

Дані заімпортовано за допомогою функції read.cvs2().

Після імпорту даних з CSV-файлу flats.csv було виконано їх статистичний опис, за допомогою функції describe() з бібліотеки psych у Rstudio (див. рис. 3.3).

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
rooms	1	216	2.01	0.97	2	1.94	1.48	1	6	5	0.73	0.44	0.07
location*	2	217	1.27	0.44	1	1.21	0.00	1	2	1	1.04	-0.91	0.03
condition*	3	217	1.77	0.42	2	1.84	0.00	1	2	1	-1.30	-0.30	0.03
m2	4	217	76.33	38.02	67	70.94	28.17	21	280	259	1.77	4.61	2.58
type*	5	216	1.20	0.40	1	1.13	0.00	1	2	1	1.46	0.14	0.03
price	6	217	82427.45	82183.66	59548	67365.84	35609.09	1	750000	749999	4.58	29.38	5578.99

Рисунок 3.3 – Описова статистика по змінним

Під час статистичного опису було розраховано основні показники для кожної змінної, що дозволяють проаналізувати їх розподіл та варіативність. Ці показники включають середнє значення, медіану, стандартне відхилення та інші параметри, які допомагають виявити особливості даних та можливі аномалії [30].

Після завершення етапу імпорту даних та було здійснено детальний аналіз статистичного опису змінних, які планується використовувати в моделі прогнозування цін на нерухомість. Кожна з цих змінних відображає ключові характеристики об’єктів нерухомості, що можуть суттєво впливати на їхню ринкову вартість. Зокрема, важливо врахувати, що аналіз надає змогу виявити закономірності, а також оцінити ступінь впливу різних параметрів на формування цін. На основі проведеного статистичного опису можна зробити висновок, що

перед використанням даних у моделі прогнозування необхідно виконати додаткову обробку.

Аналіз виявив наявність пропущених значень, а також значень, що не відповідають логічним або статистичним очікуванням. Ці аномалії можуть негативно вплинути на точність та надійність моделі, оскільки вона може сприймати їх як валідні дані, що може призвести до спотворення результатів. Тому важливо реалізувати етапи очищення даних, включаючи заповнення пропусків, виправлення неадекватних значень або їх видалення, щоб забезпечити високу якість та коректність даних, які будуть використовуватися для подальшого аналізу і моделювання (див. табл. 3.1).

Таблиця 3.1 – Аналіз статистичного опису набору даних

№	Назва змінної	Статистичний опис
1	rooms	Кількість кімнат у квартирі. У вибірці представлено 216 квартир. Середня кількість кімнат становить 2.01, мінімальна – 1, а максимальна – 6.
2	location	Локація квартири, категоріальна змінна (значення 1 або 2). Усього 217 спостережень. Визначає географічне розташування житла, що може впливати на його ціну.
3	condition	Стан квартири, категоріальна змінна. У вибірці 217 спостережень, середнє значення – 1.77. Впливає на привабливість об'єкта для покупців.
4	m2	Площа квартири у квадратних метрах. Представлено 217 квартир. Середня площа – 76.33 м², мінімальна – 21 м², максимальна – 280 м². Площа є важливим фактором, що безпосередньо впливає на ціну.
5	type	Тип квартири, категоріальна змінна (наприклад, новобудова або вторинний ринок). Вибірка містить 216 спостережень. Тип квартири також впливає на ринкову вартість.
6	price	Ціна квартири (у гривнях). У вибірці 217 спостережень. Середня ціна становить 82 427.45 грн, з великим діапазоном від 1 грн до 1 750 000 грн. Це основний показник для аналізу та прогнозування.

У процесі обробки даних важливо провести фільтрацію, щоб зосередитися на спостереженнях, які є релевантними для подальшого аналізу [31]. На першому етапі з таблиці були видалені всі квартири з ціною, що перевищує 300 000. Це дозволяє усунути надмірно дорогі об'єкти, які можуть спотворити результати аналізу.

На другому етапі було проведено додаткову фільтрацію, яка залишила лише ті спостереження, де ціна квартири перевищує 10 000. Це також допомогло видалити квартири з аномально низькою вартістю, які не відповідають реаліям ринку.

Після фільтрації був проведений статистичний опис відфільтрованих даних для ключових змінних: кількість кімнат, площа квартири в квадратних метрах та ціна. Це дозволило отримати зведену статистику для цих трьох змінних після очищення набору даних.

Завдяки виконаним крокам таблиця спостережень була суттєво очищена. Кількість об'єктів у вибірці зменшилася з 217 до 213, оскільки були видалені всі спостереження, що не відповідали критеріям фільтрації (див. рис. 3.4). Це підвищує якість даних та забезпечує коректність подальшого аналізу і моделювання.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
rooms	1	212	1.98	0.94	2	1.91	1.48	1	6	5	0.71	0.46	0.06
m2	2	213	73.95	33.12	67	69.78	28.17	21	212	191	1.22	1.56	2.27
price	3	213	75524.68	52002.81	59538	66041.70	34973.05	15000	280000	265000	1.74	2.94	3563.17

Рисунок 3.4 – Описова статистика по змінним після процесу фільтрації

Для наочного аналізу змінних у наборі даних було створено гістограми, які дозволяють візуально оцінити розподіл значень кожної зі змінних. Це дає змогу швидше зрозуміти структуру даних та виявити ключові тенденції, такі як частота появи різних значень, домінуючі категорії та потенційні аномалії. Гістограми слугують ефективним інструментом для візуалізації, що полегшує інтерпретацію результатів і підготовку до подальшого аналізу.

Гістограми, ілюструють розподіл трьох ключових змінних, що характеризують об’єкти нерухомості: кількість кімнат (rooms), площу (m2) та ціну (price).

Перша гістограма відображає розподіл кількості кімнат у квартирах. З даних видно, що найбільше квартир має 2 та 3 кімнати, а також значна частина об’єктів є 1-кімнатними. У той же час, квартири з 4 і більше кімнатами зустрічаються рідше. Це свідчить про те, що найбільш популярними є невеликі та середні за розміром квартири. Відносно невеликий попит на об'єкти з великою кількістю кімнат може вказувати на специфічні потреби та уподобання споживачів на ринку нерухомості.

Друга гістограма представляє площу квартир у квадратних метрах. Більшість об'єктів має площу в межах від 50 до 100 квадратних метрів, в той час як значна кількість квартир має площу від 100 до 150 м². Однак розподіл площі стає менш рівномірним для великих розмірів, оскільки частота таких об'єктів зменшується. Це підкреслює, що ринок переважно пропонує квартири середнього розміру, які найчастіше обирають покупці.

Третя гістограма ілюструє ціновий діапазон квартир. Найбільше об'єктів має ціну в межах від 50 000 до 100 000, тоді як кількість квартир, що потрапляють у нижчу цінову категорію або перевищують 200 000, є досить обмеженою. Це свідчить про те, що більшість пропозицій на ринку нерухомості знаходяться в доступному ціновому діапазоні, а ціни на квартири значною мірою коливаються в межах середньої цінової категорії (див. рис. 3.5).

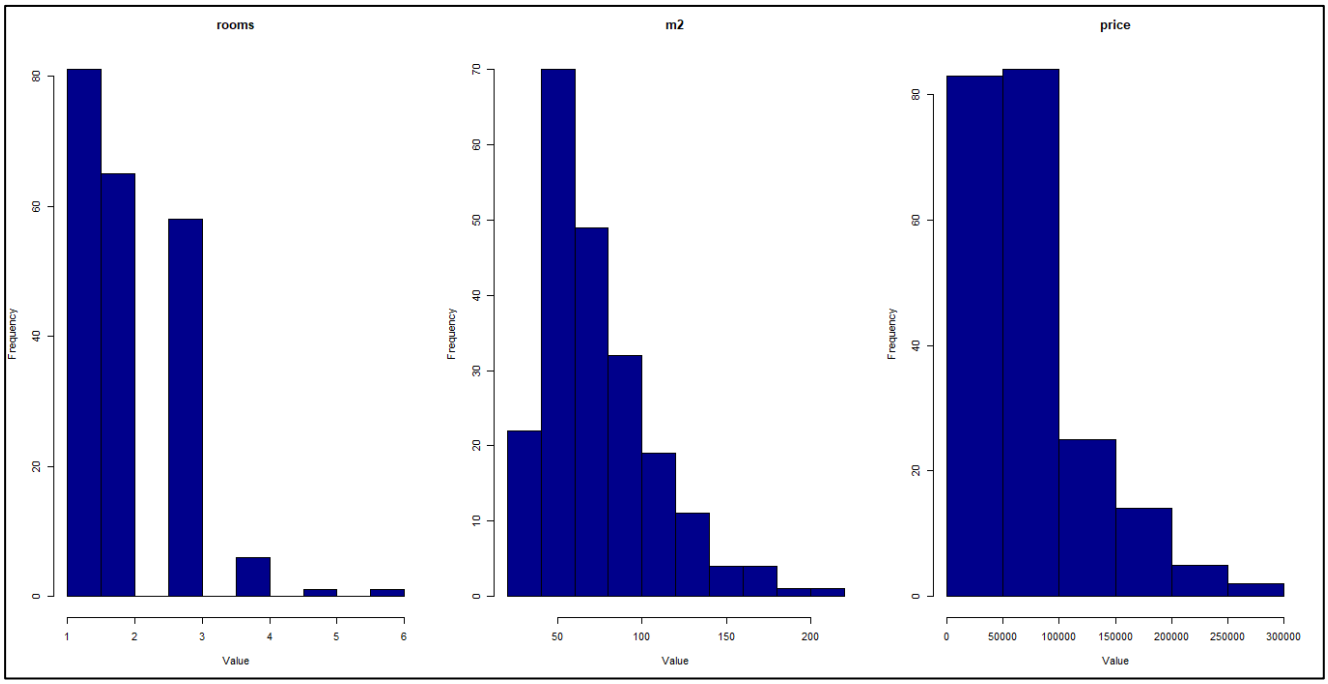

Рисунок 3.5 – Гістограми змінних rooms, m2 та price

Для забезпечення коректності та надійності аналізу даних у наборі було проведено перевірку на наявність відсутніх значень у колонках, що містять важливу інформацію, зокрема у змінних rooms (кількість кімнат) і type (тип квартири) [31].

Спочатку була виконана команда any(is.na(f)), яка дозволяє визначити, чи є в наборі даних принаймні одне відсутнє значення (NA). Якщо команда повертає TRUE, це свідчить про те, що у хоча б одній колонці присутні NA, що може негативно вплинути на результати подальшого аналізу.

Для більш детального аналізу була використана команда colSums(is.na(f[, c("rooms", "type")])). Ця команда підраховує кількість відсутніх значень у вказаних колонках rooms та type, надаючи чітке уявлення про масштаби проблеми.

В результаті проведених перевірок були виявлені відсутні значення (NA) у цих колонках, що вказує на необхідність їх подальшої обробки (див. рис. 3.6).

```
> any(is.na(f))
[1] TRUE
> colSums(is.na(f[, c("rooms", "type")]))
rooms  type
    1     1
>
```

Рисунок 3.6 – Перевірка на відсутні значення (NA) в наборі даних

Для покращення якості даних та забезпечення коректності подальшого аналізу було прийнято рішення видалити рядки з набору даних, що містять відсутні значення в конкретних колонках, зокрема в колонках rooms (кількість кімнат) і type (тип квартири). Це є важливим кроком у підготовці даних, оскільки відсутні значення можуть суттєво вплинути на результати аналізу і в подальшому, на точність моделі прогнозування.

З метою реалізації цього підходу було використано команду, яка відфільтровує набір даних, залишаючи лише ті рядки, в яких значення у вказаних колонках є непорожніми. У результаті виконання цієї команди, всі рядки, де присутні відсутні значення (NA) в колонках rooms або type, були видалені з таблиці.

В змінну f_clean, було збережено лише ті спостереження, які мають повну інформацію, що, у свою чергу, підвищує надійність даних, що використовуються для подальшого аналізу та моделювання (див. рис. 3.7). Видалення рядків з відсутніми значеннями є одним із стандартних методів обробки даних, що допомагає уникнути спотворення результатів і забезпечити їхню адекватність.

```
> f_clean <- f[!is.na(f$rooms) & !is.na(f$type), ]
> any(is.na(f_clean))
[1] FALSE
>
```

Рисунок 3.7 – Перевірка на відсутні значення (NA) в наборі після збереження даних в змінну f_clean

У результаті перевірки в наборі даних не було знайдено відсутніх значень.

Для виявлення аномальних значень у наборі даних було використано boxplot-діаграми (ящики з "вусами") для кожної числової змінної. Цей метод візуалізації є ефективним інструментом для дослідження розподілу даних та виявлення потенційних викидів, що можуть свідчити про аномалії або помилки у вибірці. Boxplot дозволяє наочно оцінити, як розподілені значення змінних, виявити їх медіану, міжквартильний розмах, а також можливі аномалії, представлені точками, розташованими за межами "вусів".

Аналіз побудованих діаграм розмаху для числових змінних показав наявність аномальних значень у змінних m² (площа квартири) та price (ціна квартири) (див. рис. 3.8). Виявлені аномалії можуть вказувати на можливі помилки введення даних, екстремальні значення, або ж бути результатом рідкісних, але реальних випадків, таких як квартири з незвично великою площею чи ціною [32].

Значення, що знаходяться далеко за межами меж "вусів", класифікуються як викиди. Такі дані потребують додаткової перевірки та обробки, оскільки вони можуть вплинути на точність побудованої моделі.

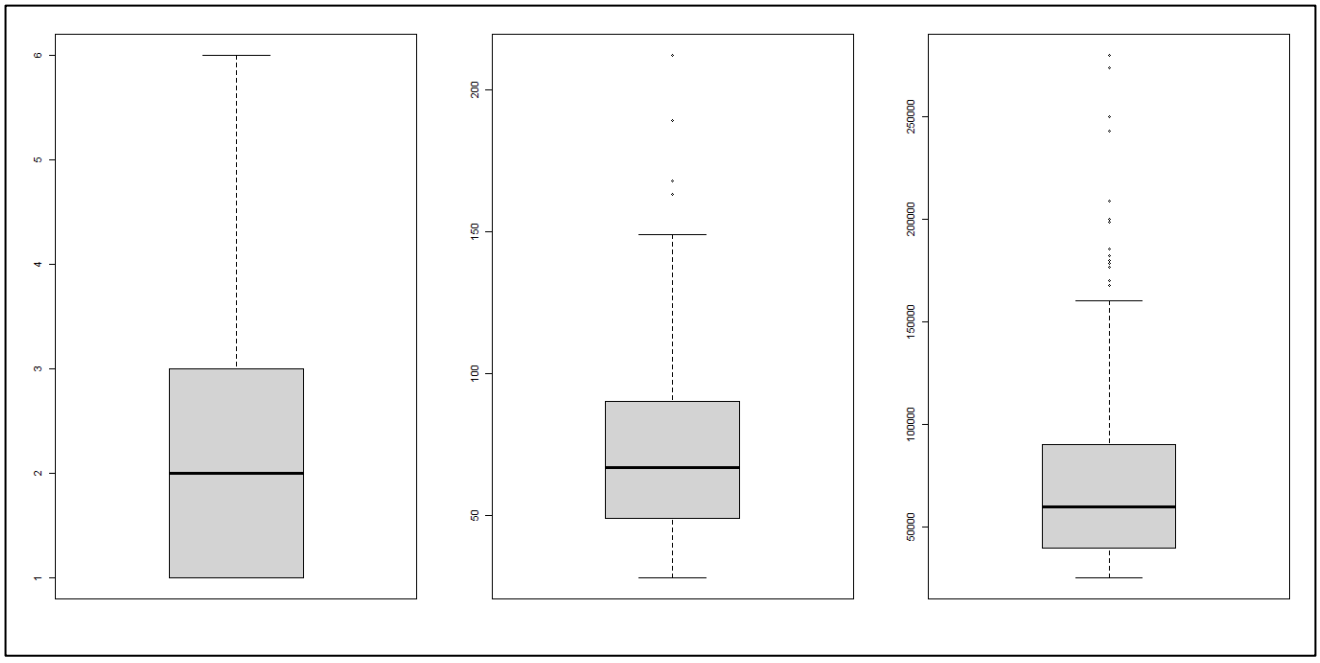

Рисунок 3.8 – Boxplot-діаграми для змінних rooms, m2 та price

Для візуалізації та подальшого аналізу впливу стану житла на його ціну було побудовано boxplot-діаграму, яка відображає розподіл цін для двох категорій: відремонтованих (repaired) та невідремонтованих (unrepaired) квартир (див. рис. 3.9).

Використання такого типу графіків є доцільним, оскільки він дозволяє побачити основні характеристики розподілу – медіану, міжквартильний розмах, а також аномальні значення, що виходять за межі типового діапазону. Такий підхід сприяє кращому розумінню впливу якісних характеристик нерухомості на цінові показники та дозволяє своєчасно виявити аномалії, що можуть викривляти загальні результати моделі.

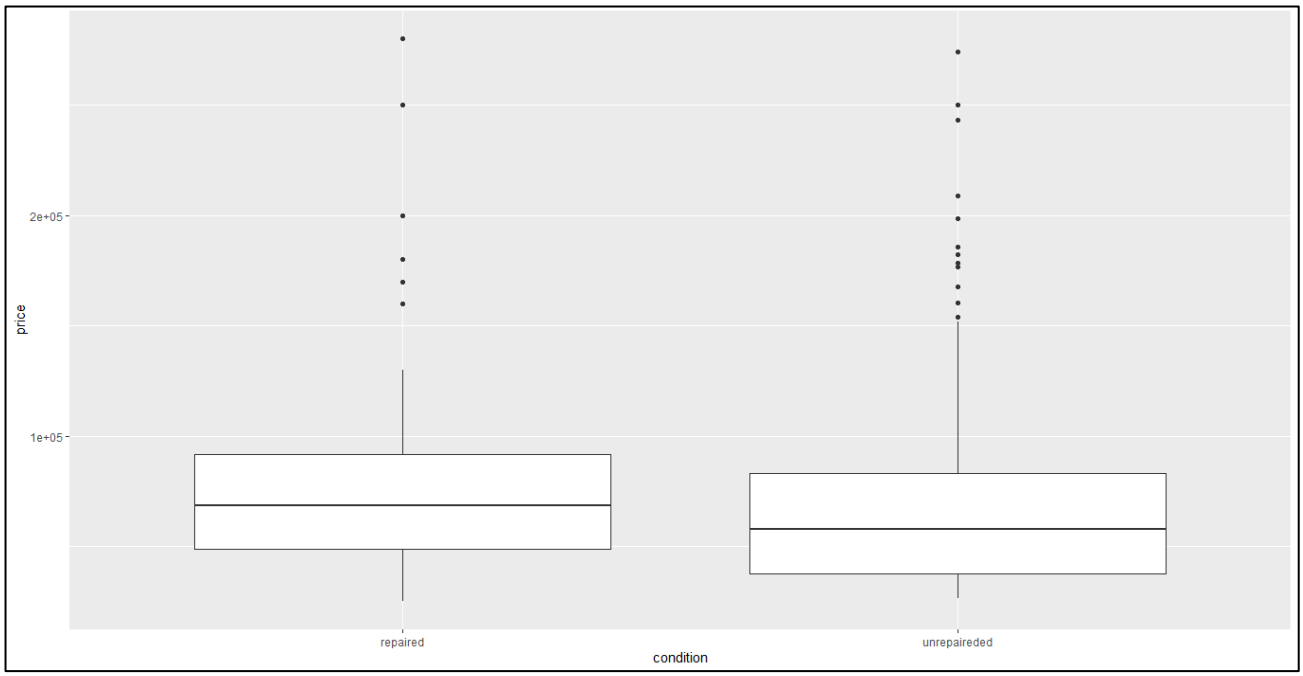

Рисунок 3.9 – Boxplot-діаграма яка відображає розподіл цін для двох категорій квартир

На діаграмі зображено, що ціни на квартири в обох категоріях мають подібні розподіли. Медіана для обох груп знаходиться приблизно на однаковому рівні, що свідчить про відсутність значної різниці у середніх цінах між відремонтованими та

невідремонтованими квартирами. Іншими словами, наявність ремонту не є вирішальним фактором у ціновому формуванні для даного вибіркового набору.

Проте, варто звернути увагу на аномальні значення, що відображені точками вище «вусів» діаграми. Ці значення свідчать про наявність квартир із цінами, що суттєво перевищують типовий діапазон для обох категорій. Для невідремонтованих квартир кількість таких аномальних точок є дещо більшою, що може вказувати на специфічні випадки, де відсутність ремонту компенсується іншими перевагами, такими як розташування чи інші характеристики об’єкта.

Загалом, результати графічного аналізу свідчать про те, що стан квартири не є домінуючим фактором у формуванні ціни.

Для коректної обробки даних і підвищення якості прогнозної моделі було виконано кілька важливих етапів попередньої підготовки даних, зокрема перетворення категоріальних змінних у числові значення та обробку пропущених даних (див. рис. 3.10). Це є необхідним для забезпечення коректної роботи моделей машинного навчання, оскільки більшість алгоритмів вимагають числові вхідні дані та не допускають пропущених значень у наборах даних.

```
## Preprocessing
# Factors as numeric
f_clean$location <- as.numeric(as.factor(f_clean$location)) - 1
f_clean$condition <- as.numeric(as.factor(f_clean$condition)) - 1
f_clean$type <- as.numeric(as.factor(f_clean$type)) - 1

# Missing data
f_clean$rooms <- ifelse(is.na(f_clean$rooms),
                  round(mean(f_clean$rooms, na.rm = TRUE)), f_clean$rooms)
f_clean$type <- ifelse(is.na(f_clean$type),
                  round(mean(f_clean$type, na.rm = TRUE)), f_clean$type)
```

Рисунок 3.10 – Фрагмент коду для перетворення категоріальних змінних у числові значення та обробки пропущених даних

У перших рядках коду категоріальні змінні location, condition, і type були перетворені на числові значення. Це зроблено за допомогою функцій as.factor() та

as.numeric(). Кожне унікальне значення змінної було закодовано числом, що дозволяє використовувати ці змінні у машинному навчанні та аналітичних моделях.

Друга частина коду присвячена роботі з пропущеними значеннями у колонках rooms (кількість кімнат) та type (тип нерухомості). У цих випадках пропущені значення (NA) замінюються на середнє значення відповідної змінної. Використовується функція ifelse() для перевірки наявності пропущених значень: якщо вони присутні, у відповідну позицію підставляється округлене середнє значення змінної.

Причини для використання даного підходу:

– уникнення втрати даних – заміна пропущених значень середніми дозволяє зберегти рядки, які містять відсутні дані, і уникнути їх видалення;

– стабільність моделі – використання середнього значення зменшує вплив аномальних даних порівняно з видаленням рядків;

– коректність результатів – цей підхід допомагає підтримувати загальний розподіл даних, уникаючи спотворень через видалення пропущених значень.

Даний етап підготовки є критично важливим для подальшого аналізу та побудови моделей. Перетворення категоріальних змінних у числові забезпечує сумісність з методами машинного навчання [32]. Обробка пропущених значень шляхом заміни на середні значення дозволяє уникнути проблем із недостатнім обсягом даних та зберегти цінну інформацію.

Для всебічного аналізу розподілу змінних після логарифмічного перетворення було побудовано діаграми, що ілюструють зміни у характеристиках досліджуваних даних (див. рис. 3.11).

```
## Visualising
library(ggplot2)
par(mfrow = c(2, 3))

hist(f_clean$rooms, col = 'dark blue', main = 'rooms', xlab = 'value')
hist(f_clean$m2, col = 'dark blue', main = 'm2', xlab = 'value')
hist(f_clean$price, col = 'dark blue', main = 'price', xlab = 'value')

hist(log(f_clean$rooms), col = 'dark blue', main = 'rooms', xlab = 'value')
hist(log(f_clean$m2), col = 'dark blue', main = 'm2', xlab = 'value')
hist(log(f_clean$price), col = 'dark blue', main = 'price', xlab = 'value')
```

Рисунок 3.11 – Фрагмент коду для побудови гістограм звичайних та прологарифмованих даних

Після логарифмічного перетворення розподіл даних значно змінюється. Гістограма для кількості кімнат стає більш рівномірною, що знижує скидання та дозволяє краще порівнювати групи. Логарифмічне перетворення сприяє нормалізації розподілу, що полегшує виявлення зв'язків та тенденцій у даних. Аналогічні зміни спостерігаються і в гістограмі площі: після трансформації розподіл виглядає більш симетрично, що свідчить про те, що варіації в розмірах можуть бути більш рівномірно представлені серед різних об'єктів нерухомості.

У свою чергу, гістограма для ціни також демонструє більш симетричний вигляд після трансформації, що полегшує аналіз і інтерпретацію зв'язків між ціною та іншими факторами (див. рис. 3.12).

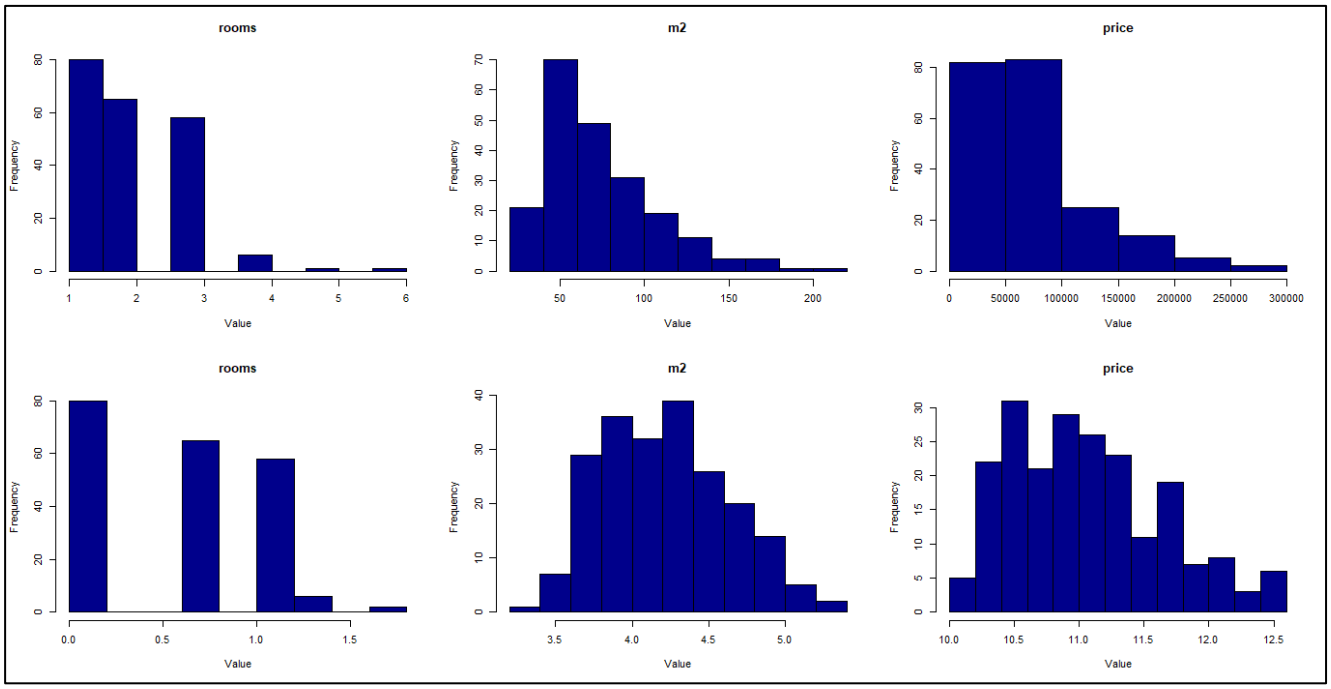

Рисунок 3.12 – Гістограми для кожної числової змінної та гістограми прологарифмованих значень кожної змінної

Логарифмічне перетворення є важливим інструментом в аналізі даних, оскільки дозволяє стабілізувати дисперсію та нормалізувати розподіли. Це робить дані більш придатними для статистичного аналізу, особливо для лінійного моделювання. Перетворення також полегшує інтерпретацію результатів.

Крім того, перетворення зменшує вплив екстремальних значень або викидів, які можуть спотворювати результати статистичних аналізів. Це особливо корисно у випадках, пов'язаних з даними про нерухомість, де об'єкти високого класу можуть непропорційно впливати на середні розрахунки.

У наступному етапі аналізу даних було здійснено логарифмічне перетворення змінних, що дозволило зменшити вплив правостороннього скидання та стабілізувати дисперсію. Цей процес був реалізований за допомогою простого коду, в якому для кожної з ключових змінних (кількість кімнат, площа в квадратних метрах та ціна) застосовувалася функція логарифма.

Результати описової статистики для трансформованих змінних показали, що середнє значення для кількості кімнат склало 0.57, із стандартним відхиленням 0.49, вказуючи на невелику варіацію та симетричність розподілу.

Для площі (м²) середнє значення становить 4.22, зі стандартним відхиленням 0.41, що свідчить про помірне скидання. Щодо ціни, середнє значення склало 11.05, з стандартним відхиленням 0.58, вказуючи на легке скидання (див. рис. 3.13).

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
rooms	1	211	0.57	0.49	0.69	0.56	0.60	0.00	1.79	1.79	0.00	-1.39	0.03
m2	2	211	4.22	0.41	4.20	4.20	0.45	3.33	5.36	2.02	0.30	-0.56	0.03
price	3	211	11.05	0.58	10.99	11.01	0.61	10.13	12.54	2.42	0.52	-0.49	0.04

Рисунок 3.13 – Описова статистика по змінним після процесу логарифмування

Логарифмування значно допомогло нормалізувати розподіл даних, що є корисним для подальшого аналізу та побудови моделей. Тепер розподіли є менш асиметричними і краще підходять для використання в регресійних та інших моделях.

У наступному етапі аналізу даних було реалізовано код, спрямований на обробку викидів у змінних, що досліджуються: кількість кімнат, площа (м²) та ціна. Використання методу ifelse дозволило виявити та замінити значення, що виходять за межі трьох стандартних відхилень від середнього значення. Така процедура є важливою для покращення якості даних, оскільки викиди можуть спотворювати результати аналізу та моделювання.

Код перевіряє, чи кожне значення у змінній перевищує три стандартних відхилення від середнього. Якщо це так, то таке значення замінюється на верхню межу, визначену як середнє значення плюс три стандартних відхилення.

Аналогічно, якщо значення менше ніж три стандартних відхилення нижче середнього, воно замінюється на нижню межу, яка розраховується як середнє мінус три стандартних відхилення (див. рис. 3.14).

```
# Replace ejections with max (no need)
f_clean$rooms <- ifelse(f_clean$rooms < mean(f_clean$rooms) + sd(f_clean$rooms) * 3,
                        f_clean$rooms, mean(f_clean$rooms) + sd(f_clean$rooms) * 3)
f_clean$rooms <- ifelse(f_clean$rooms > mean(f_clean$rooms) - sd(f_clean$rooms) * 3,
                        f_clean$rooms, mean(f_clean$rooms) - sd(f_clean$rooms) * 3)

f_clean$price <- ifelse(f_clean$price < mean(f_clean$price) + sd(f_clean$price) * 3,
                        f_clean$price, mean(f_clean$price) + sd(f_clean$price) * 3)
f_clean$price <- ifelse(f_clean$price > mean(f_clean$price) - sd(f_clean$price) * 3,
                        f_clean$price, mean(f_clean$price) - sd(f_clean$price) * 3)

f_clean$m2 <- ifelse(f_clean$m2 < mean(f_clean$m2) + sd(f_clean$m2) * 3,
                     f_clean$m2, mean(f_clean$m2) + sd(f_clean$m2) * 3)
f_clean$m2 <- ifelse(f_clean$m2 > mean(f_clean$m2) - sd(f_clean$m2) * 3,
                     f_clean$m2, mean(f_clean$m2) - sd(f_clean$m2) * 3)

describe(f_clean[, c('rooms', 'm2', 'price')])
```

Рисунок 3.14 – Фрагмент коду для обробки викидів

Результати описової статистики для змінних після обробки викидів показали, що середні значення та стандартні відхилення залишилися на рівні 0.57 для кількості кімнат, 4.22 для площі (м²) та 11.05 для ціни. Ці результати свідчать про те, що після обробки викидів дані стали більш стабільними та менш схильними до впливу екстремальних значень. Зокрема, зменшився вплив викидів на показники скидання, що свідчить про зростання симетрії розподілу (див. рис. 3.15).

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
rooms	1	211	0.57	0.49	0.69	0.56	0.60	0.00	1.79	1.79	0.00	-1.39	0.03
m2	2	211	4.22	0.41	4.20	4.20	0.45	3.33	5.36	2.02	0.30	-0.56	0.03
price	3	211	11.05	0.58	10.99	11.01	0.61	10.13	12.54	2.42	0.52	-0.49	0.04

Рисунок 3.15 – Описова статистика по змінним після процесу обробки викидів

Для з'ясування наявності та характеру зв'язку незалежних змінних із залежною змінною та один з одним побудована матриця кореляції. Вона показує кореляцію кожної пари числових змінних з набору даних (див. рис. 3.16).

	rooms	m2	price
rooms	1.0000000	0.7403050	0.6260738
m2	0.7403050	1.0000000	0.8891463
price	0.6260738	0.8891463	1.0000000

Рисунок 3.16 – Матриця кореляції

Матриця кореляції показує взаємозв'язок між трьома змінними: кількість кімнат (rooms), площа квартири (m2), та ціна квартири (price). Кореляційні значення коливаються в межах від -1 до 1, де:

- **1** означає ідеальну позитивну кореляцію (збільшення однієї змінної призводить до пропорційного збільшення іншої);
- **-1** означає ідеальну негативну кореляцію (збільшення однієї змінної призводить до зменшення іншої);
- **0** вказує на відсутність кореляції.

Згідно з отриманими результатами, спостерігаємо, що кореляційний коефіцієнт між кількістю кімнат та площею (м²) становить 0.74, що свідчить про сильний позитивний зв'язок. Це означає, що в більшості випадків з підвищенням кількості кімнат спостерігається також збільшення площі квартири. Така закономірність є логічною, оскільки зазвичай більші квартири мають більше кімнат.

Кореляційний коефіцієнт між кількістю кімнат та ціною становить 0.63, що також вказує на позитивний зв'язок між цими змінними, хоча й трохи слабший, ніж між кількістю кімнат та площею. Це свідчить про те, що квартири з більшою кількістю кімнат мають тенденцію до вищої ціни, але на ціну можуть впливати й інші фактори, такі як розташування, стан нерухомості тощо.

Кореляційний коефіцієнт між площею (м²) та ціною становить 0.89, що вказує на дуже сильний позитивний зв'язок. Це підтверджує те, що більші квартири, як правило, коштують дорожче, що є очікуваним у контексті ринку нерухомості.

Таким чином, отримані результати кореляційного аналізу демонструють значні взаємозв'язки між досліджуваними змінними, що може слугувати підґрунтям для подальшого аналізу та моделювання на ринку нерухомості.

Візуалізуємо відношення між числовими змінними за допомогою діаграм розсіювання у вигляді розширеної матриці розсіювання (див. рис. 3.17).

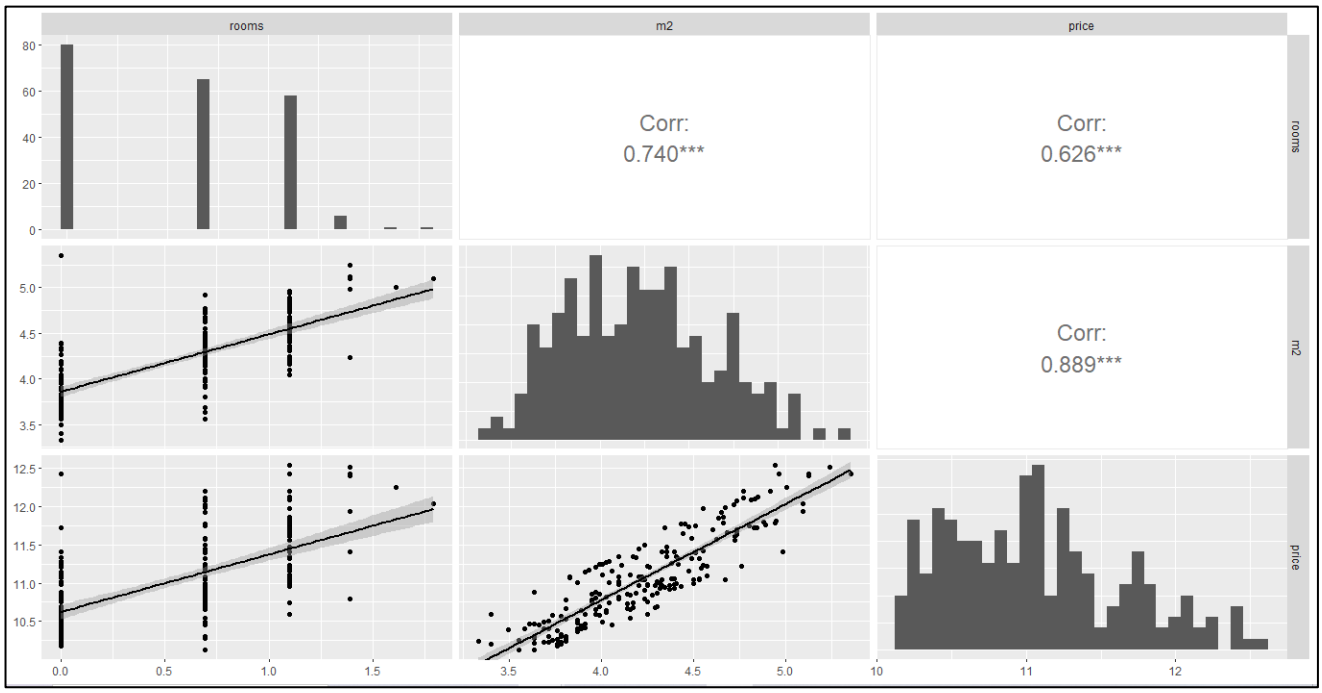

Рисунок 3.17 – Розширена матриця розсіювання для змінних rooms, m2, та price

Ця матриця, що поєднує в собі гістограми, графіки розсіювання та лінії тренду, надає можливість більш детального аналізу кореляцій між змінними.

Гістограми на діагоналі ілюструють розподіл кожної змінної. Кількість кімнат демонструє скупчення значень на нижніх рівнях, зокрема 1-3 кімнати, що підтверджує інформацію про те, що більшість квартир мають невелику кількість кімнат. Для площі (м²) гістограма показує, що більшість об'єктів нерухомості розташовані в середньому діапазоні. Гістограма для ціни вказує на зосередження даних у нижчому ціновому сегменті, проте також видно присутність вищих цін, що може свідчити про наявність деяких більш дорогих об'єктів.

На нижніх частинах матриці представлені графіки розсіювання, які відображають взаємозв'язки між змінними. Наприклад, графік, що зображає залежність площі (м²) від кількості кімнат, демонструє позитивну кореляцію з лінією тренду, що підтверджує, що квартири з більшою кількістю кімнат зазвичай мають більшу площу. Аналогічно, графік, що відображає зв'язок між ціною і площею, також має позитивну кореляцію, де видно, що зростання площі супроводжується підвищенням ціни.

Кореляційні коефіцієнти, наведені в матриці, підтверджують вищезазначені спостереження: кореляція між кількістю кімнат і площею становить 0.74, між кількістю кімнат та ціною – 0.63, а між площею та ціною – 0.89. Ці показники вказують на сильні позитивні зв'язки, які свідчать про те, що ці фактори істотно впливають один на одного в контексті ринку нерухомості.

Останнім кроком попередньої обробки даних стало розподіл очищеного набору даних на дві частини: навчальну та тестову вибірки (див. рис. 3.18). Це є важливим етапом у підготовці даних для подальшого аналізу та моделювання, оскільки дозволяє забезпечити об'єктивну оцінку якості моделей, які будуть побудовані на основі цих даних.

```
## Splitting the data set into the TRAIN set and TEST set
set.seed(123)  # Встановлюємо фіксоване значення для генератора випадкових чисел, щоб результати були відтворюваними
library(caTools)  # Завантажуємо бібліотеку caTools для функції split

split = sample.split(f_clean$price, SplitRatio = 0.8)  # Ділимо дані. 80% йде у навчальну вибірку, 20% – у тестову
f_train = subset(f_clean, split == TRUE)  # Створюємо навчальну вибірку
f_test = subset(f_clean, split == FALSE)  # Створюємо тестову вибірку

# Записуємо підготовлені дані у файли
write.csv2(f_train, file = "flats_train.csv")
write.csv2(f_test, file = "flats_test.csv")
```

Рисунок 3.18 – Код для розділення даних на навчальну та тестову вибірки

Згідно з прийнятою практикою, дані були розділені у співвідношенні 80:20. Таким чином, 80% даних було відведено для навчальної вибірки, що використовуватиметься для тренування моделей і виявлення закономірностей, тоді як 20% залишилися для тестової вибірки, призначеної для перевірки точності та

надійності отриманих моделей. Це дозволяє уникнути перенавчання, коли модель показує високі результати на навчальних даних, але має погану продуктивність на нових, невідомих даних.

Обидва набори даних були збережені у форматі CSV (*.csv), що є універсальним і зручним форматом для подальшої обробки та аналізу в різних програмних середовищах.

Цей крок не лише сприяє підготовці даних до аналізу, але й забезпечує основу для надійного тестування моделей, що є критично важливим для досягнення точних та обґрунтованих результатів у рамках дослідження ринку нерухомості. Розділення даних на навчальну та тестову вибірки є важливим етапом, що підкреслює серйозний підхід до аналізу та моделювання, орієнтованого на практичні застосування отриманих результатів.

Висновки до розділу 3

У третьому розділі роботи проведено аналіз та попередню обробку даних, що є ключовими етапами для створення інформаційної системи прогнозування цін на ринку нерухомості. Розпочато з імпорту даних, що містять різноманітну інформацію про характеристики квартир та їхні ціни.

По-перше, було здійснено статистичний опис даних, що дозволило виявити основні параметри, такі як середнє значення, медіана та стандартне відхилення для кожної змінної. Цей етап сприяв розумінню структури даних та ідентифікації можливих аномалій.

Далі проведено фільтрацію даних, що включала видалення записів із цінами, що значно перевищують середні значення. Це забезпечило підвищення якості набору даних та зменшення впливу викидів на результати аналізу.

Крім того, було здійснено перевірку на наявність пропущених значень, що дало змогу визначити ступінь заповненість даних. У разі виявлення пропусків було

ухвалено рішення про видалення рядків з такими значеннями для уникнення спотворення результатів.

Також виконано візуалізацію даних за допомогою різних графіків, що дозволило наочно продемонструвати взаємозв’язки між змінними та виявити аномальні значення. Це сприяло кращому розумінню факторів, що впливають на формування цін.

У підсумку, всі етапи попередньої обробки даних забезпечили високу якість та надійність інформації, що є критично важливими для подальшого моделювання та прогнозування цін на ринку нерухомості.

4 РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ ПРОГНОЗУВАННЯ ЦІН НА РИНКУ НЕРУХОМОСТІ

4.1 Побудова моделей

Для ефективного прогнозування цін на ринку нерухомості в рамках даної роботи було обрано кілька методів, які демонструють високу продуктивність у задачах регресії [33-41]:

- лінійна регресія;
- множинна регресія;
- поліноміальна регресія;
- дерева рішень;
- випадковий ліс;
- XGBoost.

Цей вибір методів забезпечить комплексний підхід до аналізу даних та підвищить точність прогнозів у сфері нерухомості.

Лінійна регресія є одним із найпопулярніших статистичних методів для моделювання залежності між змінними, зокрема для прогнозування. Цей метод передбачає встановлення лінійної залежності між незалежною змінною (фактором) та залежною змінною (результатом) [38]. У випадку нашого дослідження площа (в метрах квадратних) виступає як незалежна змінна, а ціна нерухомості – як залежна.

Коефіцієнти, отримані в результаті регресії, вказують на те, що площа об'єкта (m2) є значущим предиктором для ціни, оскільки p-значення для обох коефіцієнтів менше 0.05. Значення R-квадрат (приблизно 0.7938) свідчить про те, що близько 79.38% варіацій у ціні можуть бути пояснені варіаціями у площі. Це свідчить про хорошу якість моделі та можливість використання отриманих результатів для прогнозування цін на ринку нерухомості (див. рис. 4.1).

```
Call:
lm(formula = price ~ m2, data = f_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61991 -0.20049 -0.05783  0.18895  0.57358

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.71480    0.21299   26.83   <2e-16 ***
m2           1.26669    0.05011   25.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2672 on 166 degrees of freedom
Multiple R-squared:  0.7938,    Adjusted R-squared:  0.7926
F-statistic: 639.1 on 1 and 166 DF,  p-value: < 2.2e-16
```

Рисунок 4.1 – Результати моделі

На наступному етапі виконується прогнозування цін на нерухомість за допомогою моделі лінійної регресії. Зокрема, використовуються отримані з навчального набору дані для формування прогнозу цін на основі площі. Код застосовує модель, щоб здійснити прогнозування на основі тестового набору даних.

Після прогнозування обчислюються середні квадратичні помилки (MSE) для оцінки точності моделі (див. рис. 4.2). Вони розраховуються для навчального та тестового наборів даних.

```
> train_mse_sr
[1] 0.07056366
> test_mse_sr
[1] 0.07448806
```

Рисунок 4.2 – MSE для моделі лінійної регресії

Результати, отримані після виконання цього коду, показують, що середня квадратична помилка для навчального набору становить приблизно 0.0706, а для тестового набору – близько 0.0745. Це свідчить про те, що модель демонструє непогану точність прогнозування, зокрема значення MSE для тестового набору є

незначно вищим, що є типовим явищем у випадках, коли модель адаптується до навчальних даних. На рисунку 4.3 зображено графік, для аналізу якості лінійної регресії.

Рисунок 4.3 – Графік, для аналізу якості лінійної регресії

Графік відображає залежність між ціною нерухомості (по осі Y) та площею (по осі X) у квадратних метрах. Синя пряма лінія представляє результати лінійної регресії, яка демонструє позитивну кореляцію між площею та ціною: з збільшенням площі ціна на нерухомість, як правило, зростає.

Червоні та зелені точки на графіку представляють фактичні спостереження: червоні – дані з навчального набору, а зелені – з тестового.

Відносно щільне скупчення точок навколо регресійної лінії свідчить про те, що модель адекватно прогнозує ціну, хоча існують деякі відхилення, які можуть вказувати на присутність інших факторів, що впливають на вартість нерухомості.

Множинна лінійна регресія є потужним статистичним методом, який дозволяє моделювати залежність між кількома незалежними змінними та однією залежною змінною, що робить її корисною для прогнозування складних явищ.

Коефіцієнти, отримані в результаті множинної лінійної регресії, свідчать про те, що площа об'єкта (m2) та кількість кімнат (rooms) є значущими предикторами для ціни, оскільки p-значення для цих коефіцієнтів менше 0.05. Значення R-квадрат (приблизно 0.8235) вказує на те, що близько 82.35% варіацій у ціні можуть бути пояснені варіаціями в площі, кількості кімнат, місцезнаходженні та стані об'єкта. Це підтверджує добру якість моделі та її потенціал для точного прогнозування цін на ринку нерухомості (див. рис. 4.4).

```
Call:
lm(formula = price ~ rooms + location + condition + m2, data = f_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77906 -0.19914 -0.05066  0.21598  0.53577

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.61637    0.29967  18.742  < 2e-16 ***
rooms       -0.17087    0.06536  -2.614  0.00978 **
location    -0.14291    0.04951  -2.886  0.00443 **
condition   -0.23420    0.05424  -4.317 2.73e-05 ***
m2           1.36591    0.07781  17.554  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2495 on 163 degrees of freedom
Multiple R-squared:  0.8235,    Adjusted R-squared:  0.8192
F-statistic: 190.2 on 4 and 163 DF,  p-value: < 2.2e-16
```

Рисунок 4.4 – Результати моделі

На наступному етапі виконується прогнозування цін на нерухомість за допомогою моделі множинної лінійної регресії. Код застосовує модель, щоб здійснити прогнозування на основі тестового набору даних.

Після прогнозування обчислюються середні квадратичні помилки (MSE) для оцінки точності моделі. Вони розраховуються для навчального та тестового наборів даних (див. рис. 4.5).

```
> train_mse_opt
[1] 0.06038899
> test_mse_opt
[1] 0.04546395
```

Рисунок 4.5 – MSE для моделі множинної лінійної регресії

Результати, отримані після виконання цього коду, показують, що середня квадратична помилка для навчального набору становить приблизно 0.0604, а для тестового набору – близько 0.0455. Це свідчить про те, що модель демонструє високий рівень точності прогнозування, оскільки значення MSE для тестового набору значно менше, ніж для навчального. Така ситуація може свідчити про те, що модель добре узагальнює дані та здатна адекватно передбачати ціни на основі характеристик нерухомості. Використання множинної регресії в даному випадку виявляється ефективним інструментом для прогнозування цін на ринку нерухомості (див. рис. 4.6).

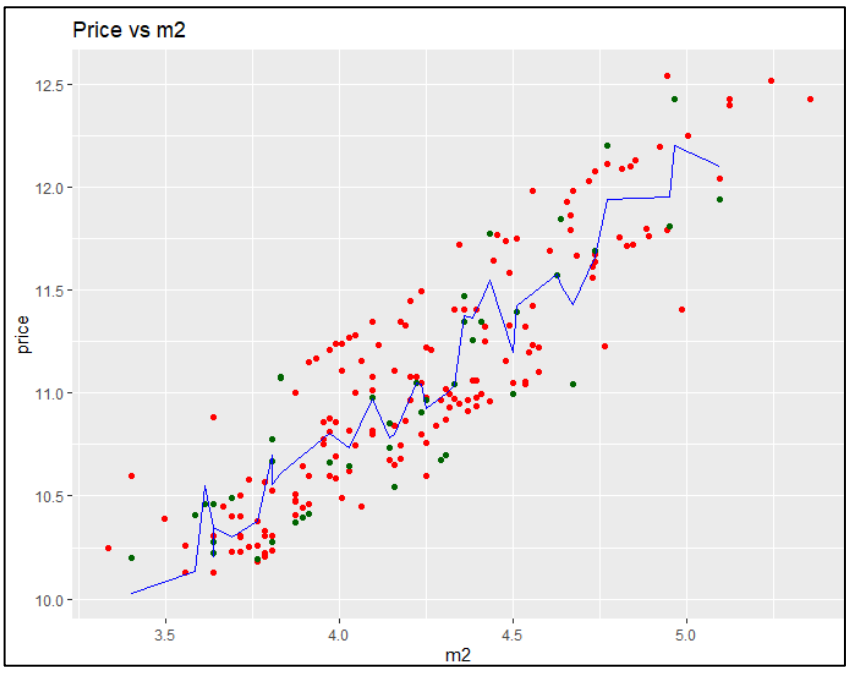

Рисунок 4.6 – Графік, для аналізу якості множинної лінійної регресії

На графіку представлено залежність ціни нерухомості від площі (м²). Сині лінії демонструють прогнозовані ціни на основі моделі, в той час як червоні та зелені точки представляють фактичні значення цін. З графіку видно, що зростання площі позитивно корелює з підвищенням ціни, оскільки загальний тренд рухається вгору. Хоча існує певна варіація у даних, модель в цілому добре відображає цю залежність, що свідчить про її адекватність для прогнозування цін на нерухомість.

Поліноміальна лінійна регресія є розширенням звичайної лінійної регресії, яке дозволяє моделювати складніші нелінійні залежності між предикторами та залежною змінною. Вона включає додавання поліноміальних членів до моделі, що забезпечує більшу гнучкість у моделюванні зв'язків між змінними (див. рис. 4.7).

```
## Polynomial Linear Regression (one factor - m2)
# Features extending
f_train_poly <- f_train[, c('price', 'm2')]
f_test_poly <- f_test[, c('price', 'm2')]

f_train_poly$m2_squared <- f_train_poly$m2^2
f_train_poly$m2_cubed <- f_train_poly$m2^3

f_test_poly$m2_squared <- f_test_poly$m2^2
f_test_poly$m2_cubed <- f_test_poly$m2^3

## 3 powers
pr <- lm(price ~ m2 + m2_squared + m2_cubed, f_train_poly)
summary(pr)
```

Рисунок 4.7 – Фрагмент коду моделі поліноміальної лінійної регресії

Результати, отримані після прогнозування, показують, що середня квадратична помилка (MSE) для навчального набору становить приблизно 0.0684, а для тестового набору – близько 0.0699. Це свідчить про те, що модель демонструє хорошу точність прогнозування, оскільки значення MSE для тестового набору є лише незначно вищим, що вказує на ефективність моделі на нових даних (див. рис. 4.8).

```
> train_mse_poly
[1] 0.06842959
> test_mse_poly
[1] 0.06994942
```

Рисунок 4.8 – MSE для моделі поліноміальної лінійної регресії

На рисунку 4.9 зображено графік , для аналізу якості поліноміальної регресії.

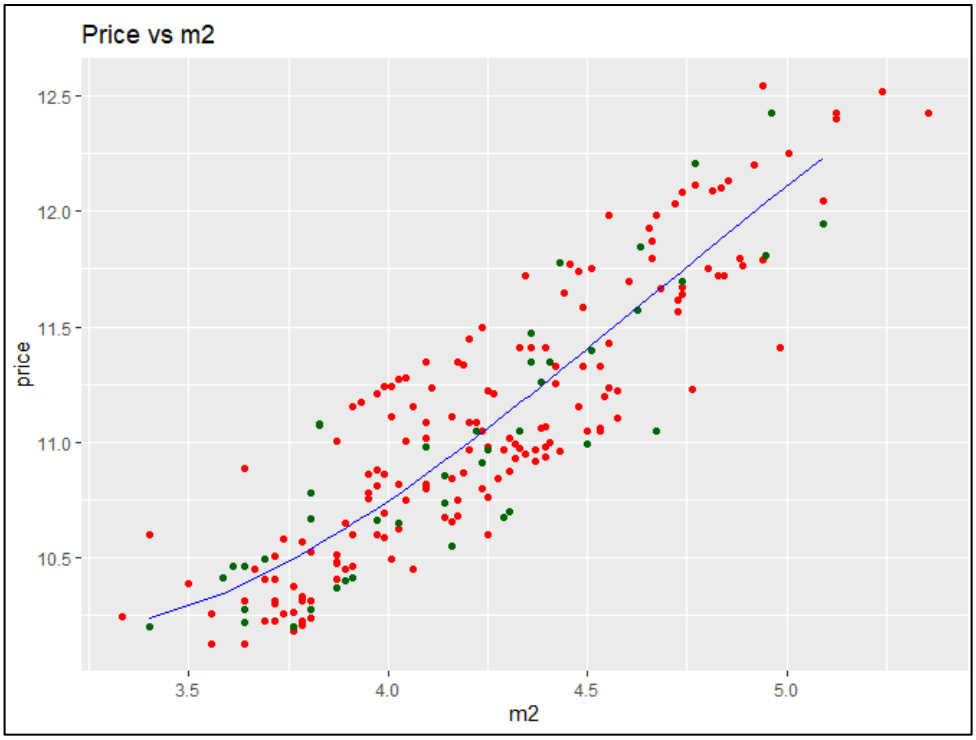

Рисунок 4.9 – Графік, для аналізу якості поліноміальної регресії

Дерева рішень є популярним методом машинного навчання, який використовується для прогнозування. Вони працюють за принципом розбиття даних на основі значень різних ознак, створюючи дерево, де кожен вузол представляє тест на певну ознаку, а кожна гілка – результат цього тесту.

Переваги цього підходу включають простоту інтерпретації та візуалізації, оскільки результати можна легко зрозуміти у формі графічної структури. Проте, незважаючи на свою простоту, дерева рішень можуть бути схильними до перенавчання, особливо при роботі з великими наборами даних, що призводить до

погіршення узагальнюючої здатності моделі на нових, невідомих даних (див. рис. 4.10).

```
> train_mse_dt
[1] 0.06702559
> test_mse_dt
[1] 0.09742735
```

Рисунок 4.10 – MSE для моделі дерева рішень

Результати, отримані після виконання цього коду, показують, що середня квадратична помилка для навчального набору становить приблизно 0.0670, а для тестового набору – близько 0.0974. Це свідчить про те, що модель демонструє задовільну точність прогнозування, хоча значення MSE для тестового набору є помітно вищим, що може вказувати на деякі проблеми з використанням даної моделі для прогнозування на наших даних. На рисунку 4.11 зображено графік, для аналізу якості дерева рішень.

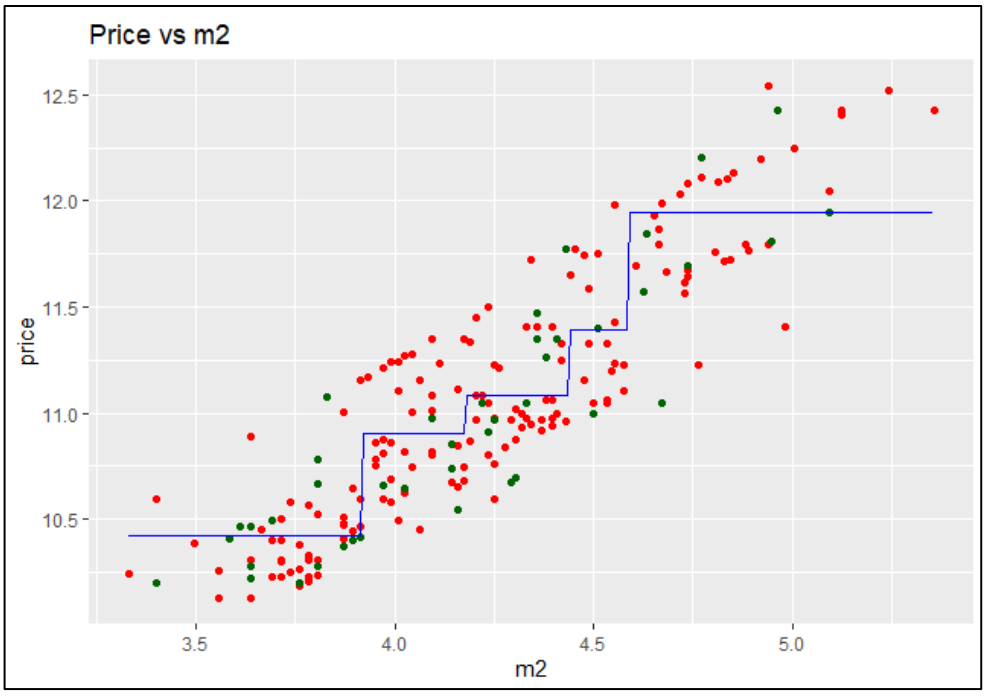

Рисунок 4.11 – Графік, для аналізу якості дерева рішень

Випадковий ліс – це потужний ансамблевий метод машинного навчання, що базується на концепції дерев рішень. Він створює безліч дерев рішень під час навчання, використовуючи різні підвибірки даних та випадкові підмножини ознак для кожного дерева, що дозволяє зменшити ризик перенавчання та підвищити стабільність моделі (див. рис. 4.12).

```
> train_mse_rf
[1] 0.04329905
> test_mse_rf
[1] 0.08267137
```

Рисунок 4.12 – MSE для моделі випадковий ліс

Отримані значення MSE, показують, що середня квадратична помилка для навчального набору становить приблизно 0.0432, а для тестового набору – близько 0.0826. Це свідчить про те, що модель демонструє задовільну точність прогнозування, хоча значення MSE для тестового набору є помітно вищим, що може вказувати на деякі проблеми з використанням даної моделі для прогнозування на наших даних. На рисунку 4.13 зображено графік, для аналізу якості моделі випадкового лісу.

XGBoost (Extreme Gradient Boosting) – це популярний метод ансамблевого навчання, який використовує технологію бустінгу для покращення результатів шляхом поетапного навчання моделей, які коригують помилки попередніх моделей. XGBoost є потужним інструментом для обробки великих наборів даних, він забезпечує високу точність прогнозів і має низку переваг, зокрема, швидкість навчання та ефективне використання пам'яті (див. рис. 4.14).

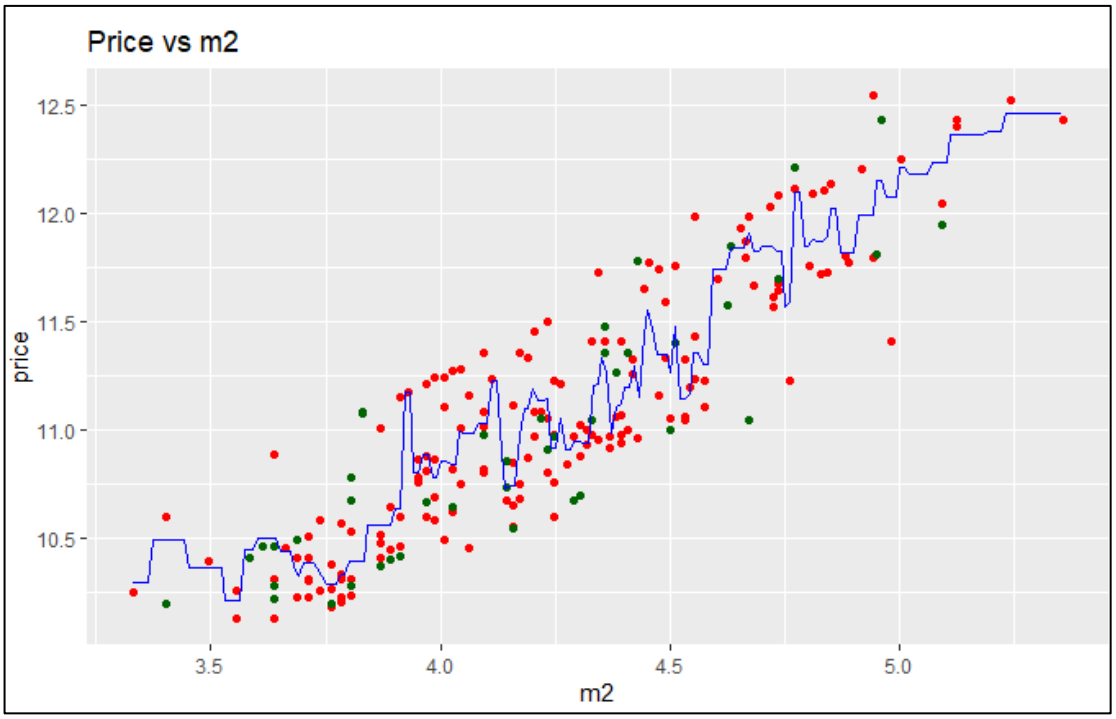

Рисунок 4.13 – Графік, для аналізу якості моделі випадковий ліс

```
# 2. Підготовка даних для XGBoost
# Перетворення даних у матрицю
dtrain <- xgb.DMatrix(data = as.matrix(f_train[, c('m2', 'rooms', 'location', 'condition')]), label = f_train$price)
dtest <- xgb.DMatrix(data = as.matrix(f_test[, c('m2', 'rooms', 'location', 'condition')]), label = f_test$price)

# 3. Налаштування параметрів для моделі XGBoost
params <- list(
  objective = "reg:squarederror", # Цільова функція для регресії
  eta = 0.1,                      # Швидкість навчання
  max_depth = 5,                  # Максимальна глибина дерева
  eval_metric = "rmse"            # Метрика для оцінки моделі
)

# 4. Навчання моделі
set.seed(1234) # Для відтворюваності результатів
nrounds <- 100 # Кількість раундів навчання
xgb_model <- xgb.train(params, dtrain, nrounds)
```

Рисунок 4.14 – Фрагмент коду для моделі XGBoost

Результати, отримані після виконання цього коду, показують, що середня квадратична помилка для навчального набору становить 0.0165, а для тестового набору – 0.0689 (див. рис. 4.15).

```
> train_mse_xgb
[1] 0.01652175
> test_mse_xgb
[1] 0.0689483
```

Рисунок 4.15 – MSE для моделі XGBoost

Це свідчить про те, що модель XGBoost демонструє дуже хорошу точність на навчальних даних, а також має деяке зниження точності на тестових даних, що є типово для моделей з високою адаптивністю до навчальних наборів (див. рис. 4.16).

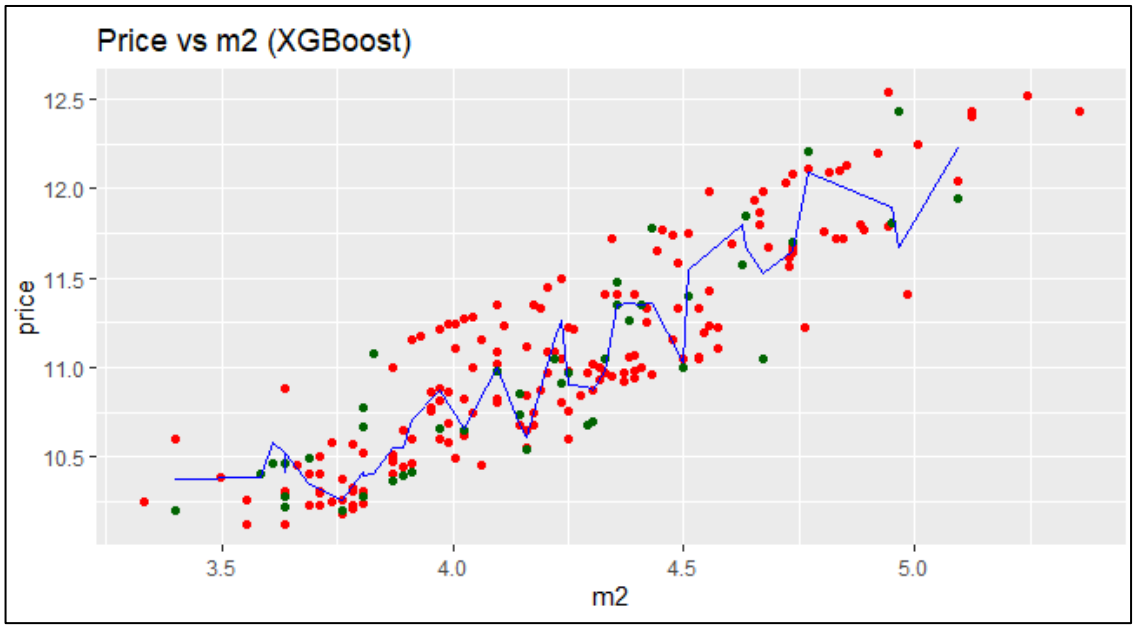

Рисунок 4.16 – Графік, для аналізу якості моделі XGBoost

4.2 Вибір найкращої моделі

Вибір найкращої моделі для інформаційної системи прогнозування цін на ринку нерухомості є важливим, оскільки від цього залежить точність і надійність прогнозів, що система надає користувачам. Правильно обрана модель дозволяє точно оцінювати вартості об'єктів, враховуючи численні фактори, що впливають на ціноутворення [42-45].

Для ефективного прогнозування цін на ринку нерухомості в рамках даної роботи було обрано найкращий метод. Для цього було проведено порівняння моделей з використанням різних критеріїв якості моделей та критеріїв якості прогнозу. До критеріїв якості моделей належать:

- коефіцієнт детермінації (R^2);
- статистика Дарбіна-Уотсона (DW);
- критерій Акаїке (AIC);
- статистика Фішера (F).

До критеріїв якості прогнозу належать:

- середня квадратична помилка (MSE);
- коефіцієнт Тейла (Theil);
- середня абсолютна відсоткова помилка (MAPE).

Ці критерії дозволяють комплексно оцінити кожну модель з різних точок зору та обрати найбільш ефективну для використання в інформаційній системі прогнозування цін на ринку нерухомості.

Коефіцієнт детермінації (R^2) показує, яку частину загальної варіації залежної змінної можна пояснити моделлю на основі незалежних змінних. Чим вищий показник R^2, тим більша здатність моделі адекватно відображати зміни в даних, що робить її більш точною та надійною для прогнозування. Моделі з високим значенням R^2 забезпечують кращу передбачуваність, оскільки вони здатні вловити більшість залежностей і мінімізувати помилки прогнозу. Найкраща модель за цим критерієм – це множинна регресія, оскільки вона показала найвищі значення R^2 серед усіх розглянутих моделей, що вказує на її високу точність та ефективність (див. рис. 4.17) .

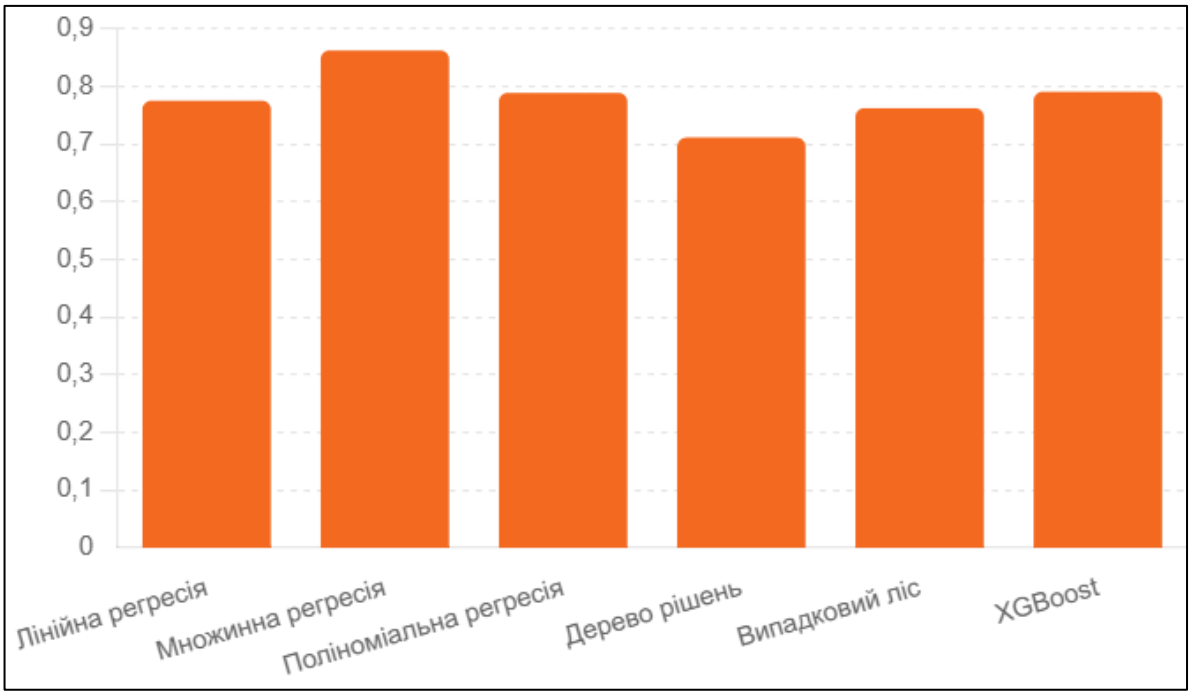

Рисунок 4.17 – Графік, якості моделей за R^2

Статистика Дарбіна-Уотсона (DW) оцінює наявність автокореляції в залишках моделі. Ідеальне значення DW наближається до 2, що свідчить про відсутність автокореляції і правильне розподілення залишків. Якщо значення DW значно відрізняється від 2, це може вказувати на проблему автокореляції, що знижує точність моделі. Найкраща модель за цим критерієм – це дерево рішень, оскільки воно продемонструвало значення DW, яке найближче до 2, що свідчить про відсутність значної автокореляції в залишках і високу стабільність моделі (див. рис. 4.18).

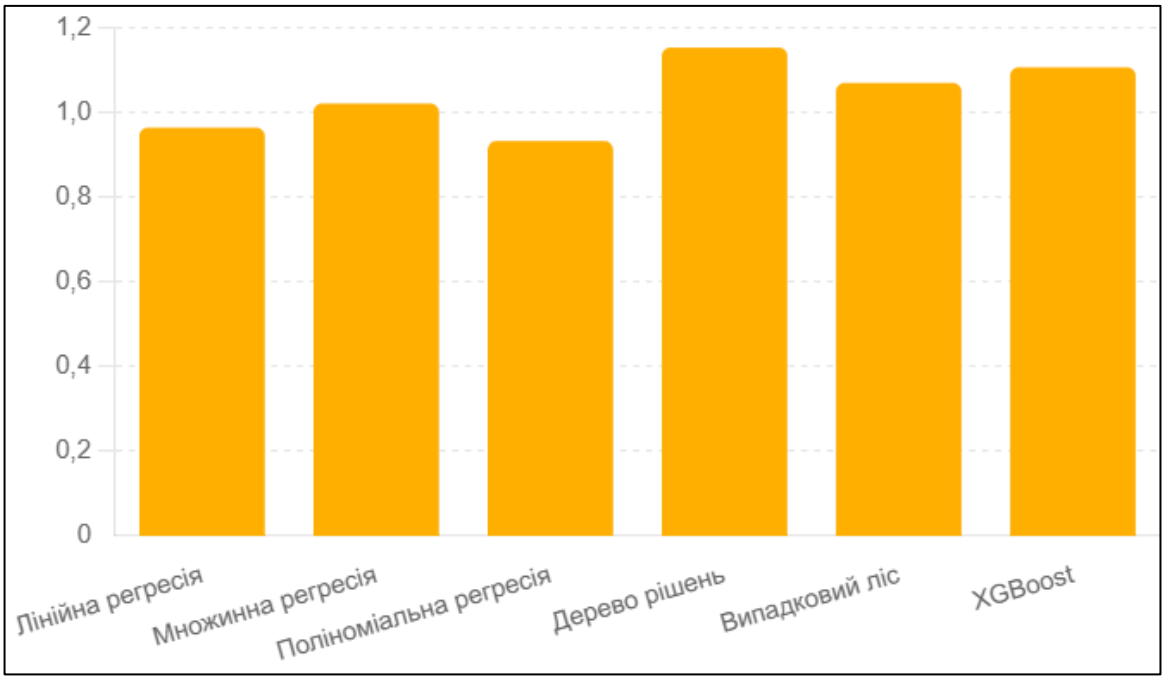

Рисунок 4.18 – Графік, якості моделей за DW

Критерій Акаїке (AIC) оцінює компроміс між точністю моделі та її складністю, штрафуючи моделі з великою кількістю параметрів. Моделі з нижчим значенням AIC є більш ефективними, оскільки вони забезпечують кращу точність без надмірної складності. Найкраща модель за цим критерієм – це множинна регресія, оскільки вона продемонструвала найнижче значення AIC серед усіх розглянутих моделей, що свідчить про її оптимальне поєднання точності та простоти (див. рис. 4.19).

Статистика Фішера (F) використовується для оцінки значущості регресійної моделі в цілому, перевіряючи, чи достатньо статистичної сили у моделі для пояснення варіацій залежної змінної. Високе значення F вказує на те, що модель здатна значно пояснити варіації в даних, що робить її більш ефективною. Найкраща модель за цим критерієм – це множинна регресія, оскільки вона показала найвище значення статистики Фішера серед усіх розглянутих моделей, що свідчить про її високу значущість та здатність точно пояснювати залежність між змінними (див. рис. 4.20).

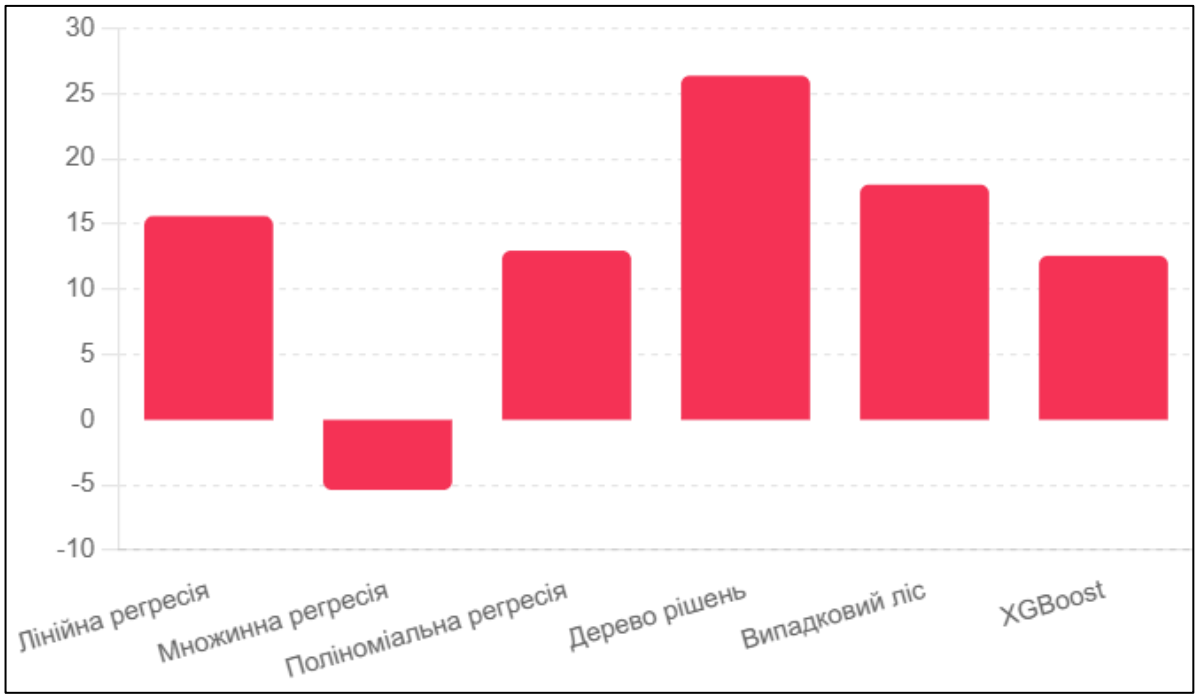

Рисунок 4.19 – Графік, якості моделей за AIC

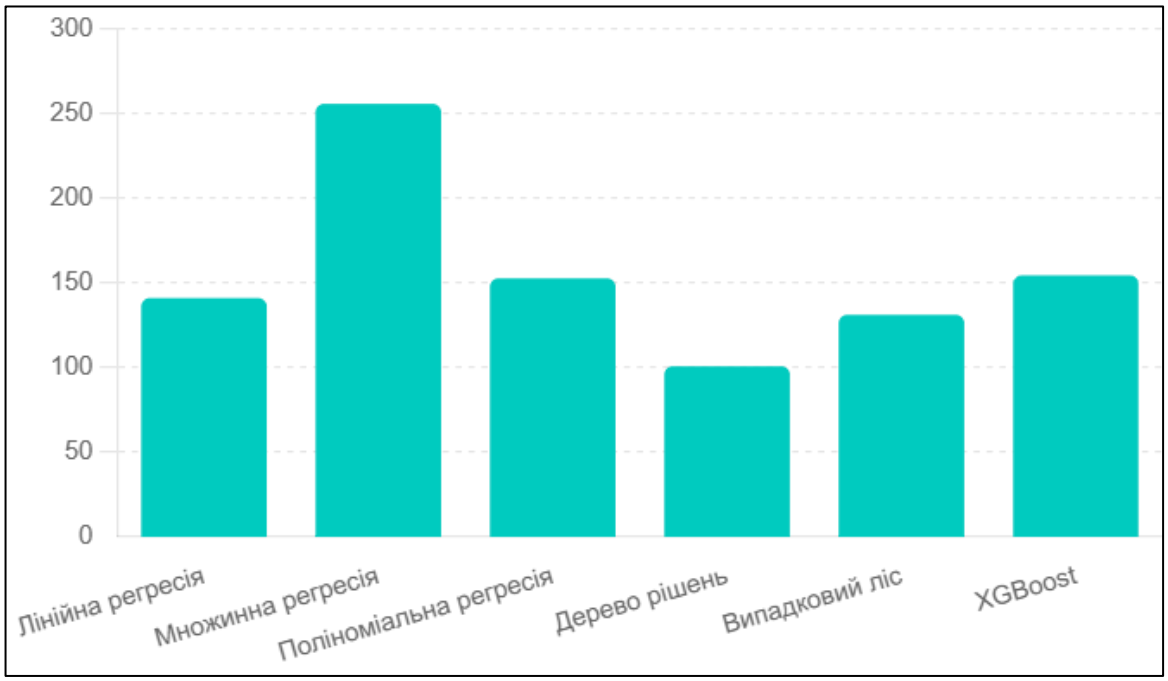

Рисунок 4.20 – Графік, якості моделей за F

Середня квадратична помилка (MSE) вимірює середнє квадратичне відхилення між прогнозованими та реальними значеннями, де менше значення

MSE свідчить про вищу точність моделі. Моделі з низьким MSE здатні мінімізувати помилки прогнозу і надають більш точні результати. Найкращий прогноз за цим критерієм у моделі множинної регресії, оскільки вона продемонструвала найнижче значення MSE серед усіх розглянутих моделей, що вказує на її високу точність у прогнозуванні цін (див. рис. 4.21).

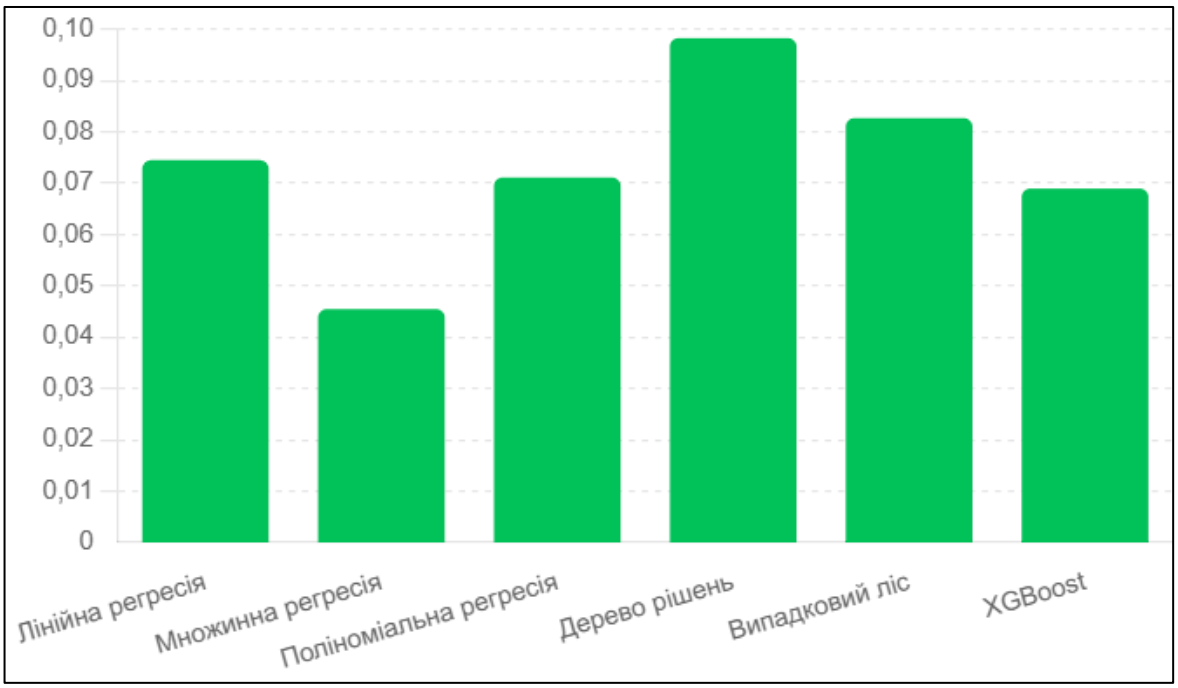

Рисунок 4.21 – Графік, якості прогнозу моделей за MSE

Коефіцієнт Тейла (Theil) оцінює точність прогнозів, порівнюючи їх з реальними значеннями на основі різних параметрів. Чим менше значення коефіцієнта Тейла, тим точніший прогноз, оскільки це вказує на менше відхилення прогнозованих значень від фактичних. Найкращий прогноз за цим критерієм у моделей множинної регресії та поліноміальної регресії, оскільки вони продемонстрували найнижче значення коефіцієнта Тейла серед усіх розглянутих моделей, що свідчить про їх високу точність у прогнозуванні цін (див. рис. 4.22).

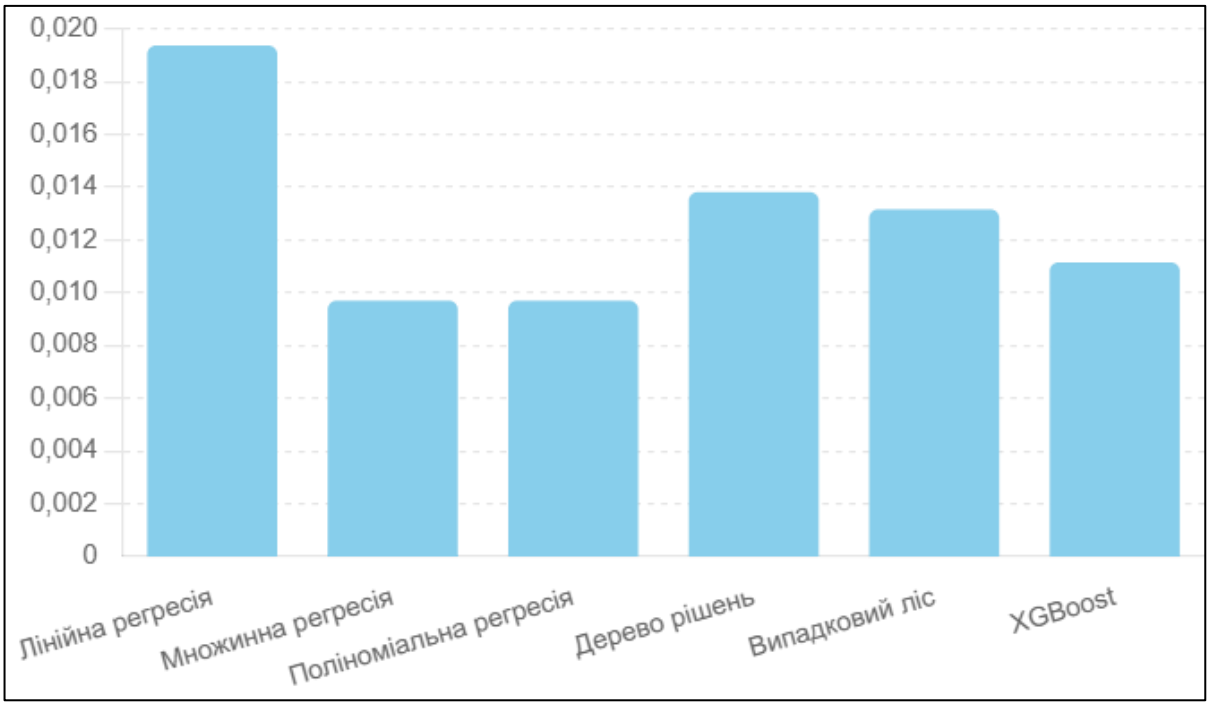

Рисунок 4.22 – Графік, якості прогнозу моделей за Theil

Середня абсолютна відсоткова помилка (MAPE) вимірює точність прогнозу в відсотках, визначаючи середнє абсолютне відхилення між прогнозованими та фактичними значеннями. Чим менше значення MAPE, тим точніший прогноз. Найкращий прогноз за цим критерієм у моделі множинної регресії, оскільки вона продемонструвала найнижче значення MAPE серед усіх розглянутих моделей (див. рис. 4.23).

На основі отриманих результатів можна зробити висновок, що **множинна регресія** є найкращим методом для прогнозування цін на нерухомість для даної системи (див. табл. 4.1). В інших моделях варто дослідити, чому вони не дають настільки хороших результатів, і можливо, провести подальшу оптимізацію або налаштування.

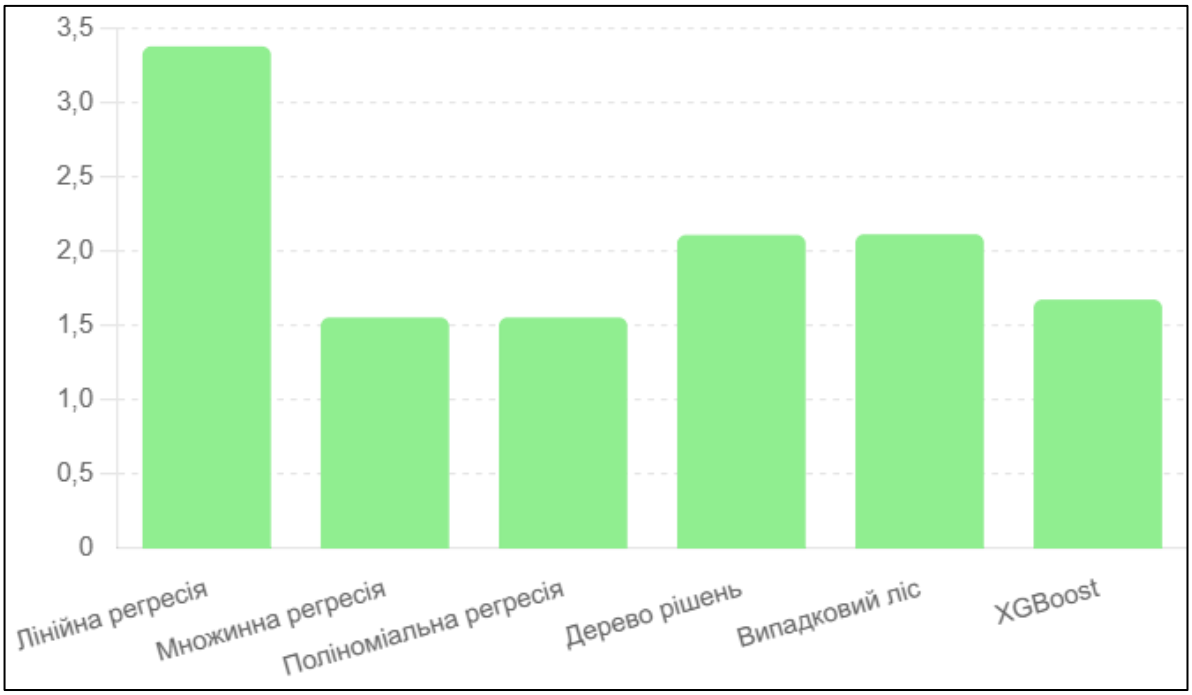

Рисунок 4.23 – Графік, якості прогнозу моделей за MAPE

Таблиця 4.1 – Узагальнена таблиця оцінки моделей

Назва моделі	**Якість моделі**				**Якість прогнозу**		
	R^2	DW	AIC	F	MSE	Theil	MAPE
Лінійна регресія	0.774	0.964	15.632	141.064	0.074	0.0193	3.379
Множинна регресія	**0.861**	1.021	**-5.378**	**255.777**	**0.045**	**0.0096**	**1.554**
Поліноміальна регресія	0.788	0.933	12.965	152.712	0.071	**0.0096**	1.555
Дерево рішень	0.71	**1.153**	26.386	100.777	0.098	0.0138	2.110
Випадковий ліс	0.761	1.07	18.03	131.188	0.082	0.0131	2.115
XGBoost	0.79	1.107	12.576	154.473	0.068	0.0111	1.675

4.3 Тестування системи

Тестування інформаційної системи прогнозування цін на ринку нерухомості проводилося для перевірки її функціональності, точності та зручності використання. Особливу увагу було приділено коректній роботі інтерфейсу програми, адже він є ключовим елементом взаємодії користувача із системою.

На початковому екрані програми розташовані всі необхідні елементи для введення параметрів об'єкта нерухомості (див. рис. 4.24).

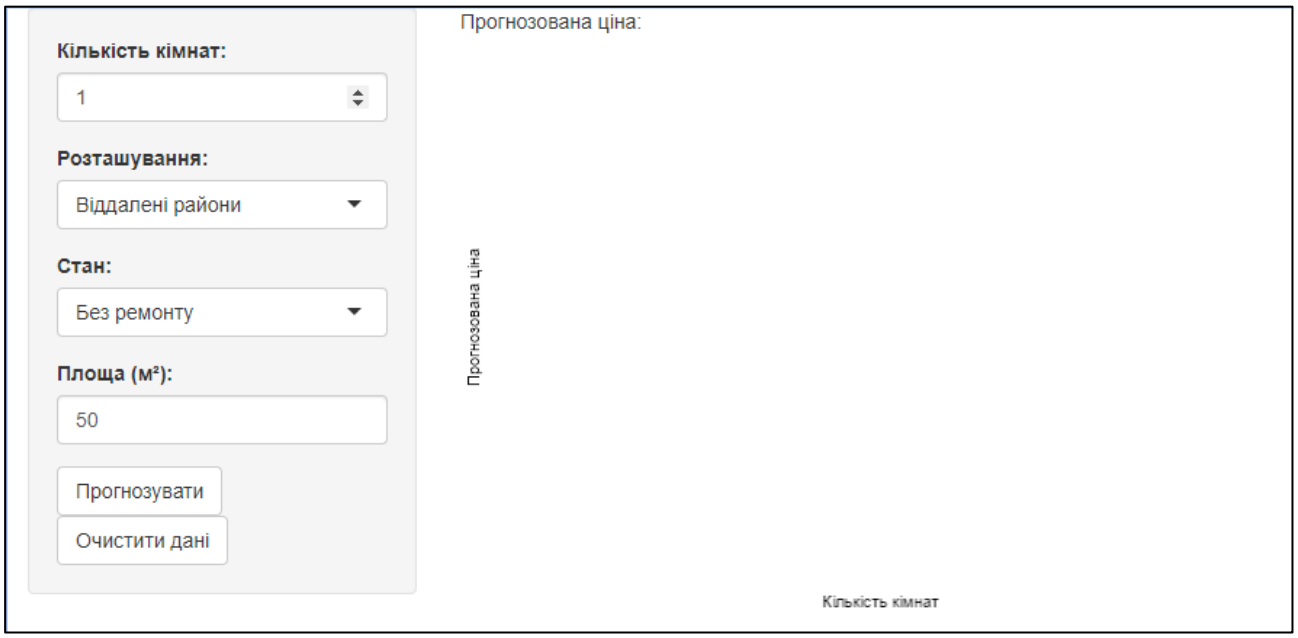

Рисунок 4.24 – Початковий екран програми

Ліва частина інтерфейсу містить панель із формами для введення ключових характеристик, які впливають на розрахунок прогнозованої ціни. Першим елементом є випадаючий список "Кількість кімнат", що дозволяє користувачу вибрати потрібну кількість від однієї і більше. Цей параметр є важливим, оскільки кількість кімнат безпосередньо впливає на ціну об'єкта.

Далі розташований випадаючий список "Розташування", у якому користувач може вибрати тип району, наприклад, "Віддалені райони". Цей пункт враховує особливості цінової політики залежно від місця розташування нерухомості, що є одним із головних факторів у формуванні вартості.

Третій параметр, "Стан", представлений випадаючим списком, де можна вказати рівень ремонту об'єкта, наприклад, "Без ремонту". Стан нерухомості суттєво впливає на її ринкову ціну, тому цей параметр був інтегрований для врахування в прогнозах.

Під вищезазначеними полями знаходиться текстове поле "Площа (м²)", у яке користувач вводить значення площі об'єкта в квадратних метрах. Це поле дозволяє вводити лише числові значення, що забезпечує коректність розрахунків.

У нижній частині панелі з елементами введення розташовані дві функціональні кнопки. Кнопка "Прогнозувати" ініціює розрахунок прогнозованої ціни на основі введених користувачем даних. Кнопка "Очистити дані" призначена для скидання всіх параметрів до початкових значень, що забезпечує зручність повторного використання програми для нових прогнозів.

Права частина екрану відведена для виведення результатів. Тут після натискання кнопки "Прогнозувати" з’являється прогнозована ціна об’єкта нерухомості, а також графічна візуалізація даних (див. рис. 4.25).

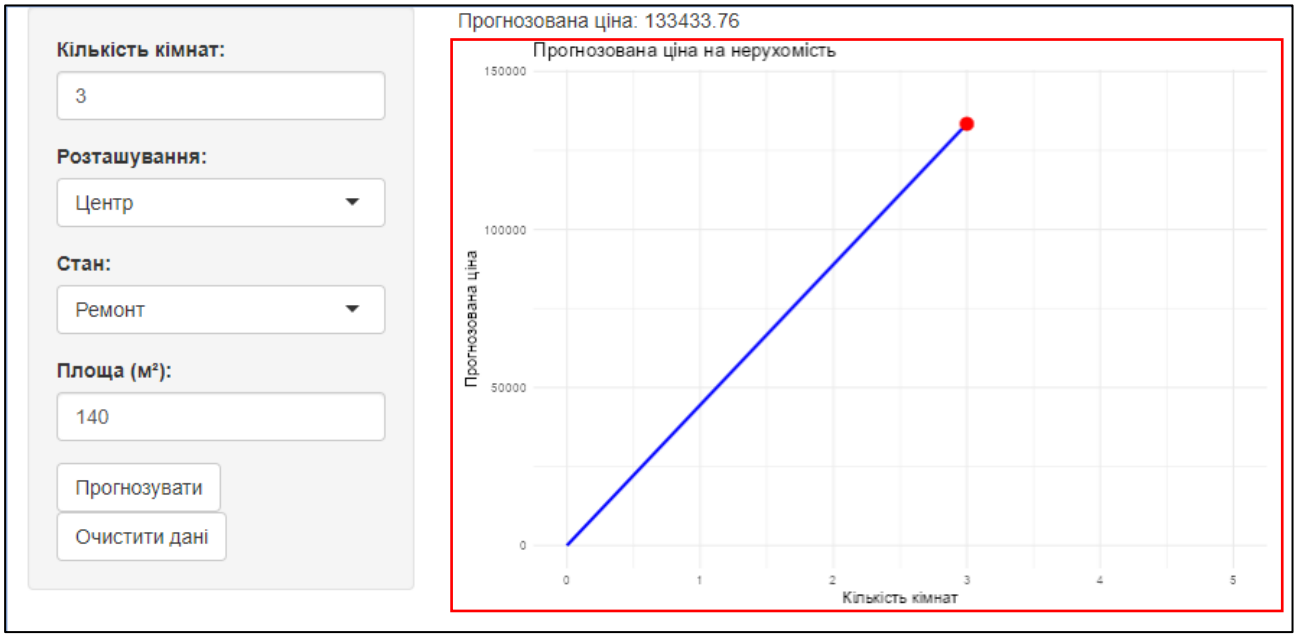

Рисунок 4.25 – Графічна візуалізація даних

Така структура інтерфейсу була спеціально розроблена для того, щоб мінімізувати час, який користувач витрачає на взаємодію з системою, і забезпечити зручність отримання результатів навіть для осіб без технічних знань.

Прогнозована ціна в системі адекватно реагує на зміну вхідних параметрів, що чітко підтверджується графічною візуалізацією результатів. У залежності від

таких характеристик, як кількість кімнат, розташування, стан ремонту та площа, система динамічно змінює прогнозовану ціну. Наприклад, зі збільшенням кількості кімнат або вибором центрального району ціна суттєво підвищується, що відповідає реальним ринковим закономірностям (див. рис. 4.26).

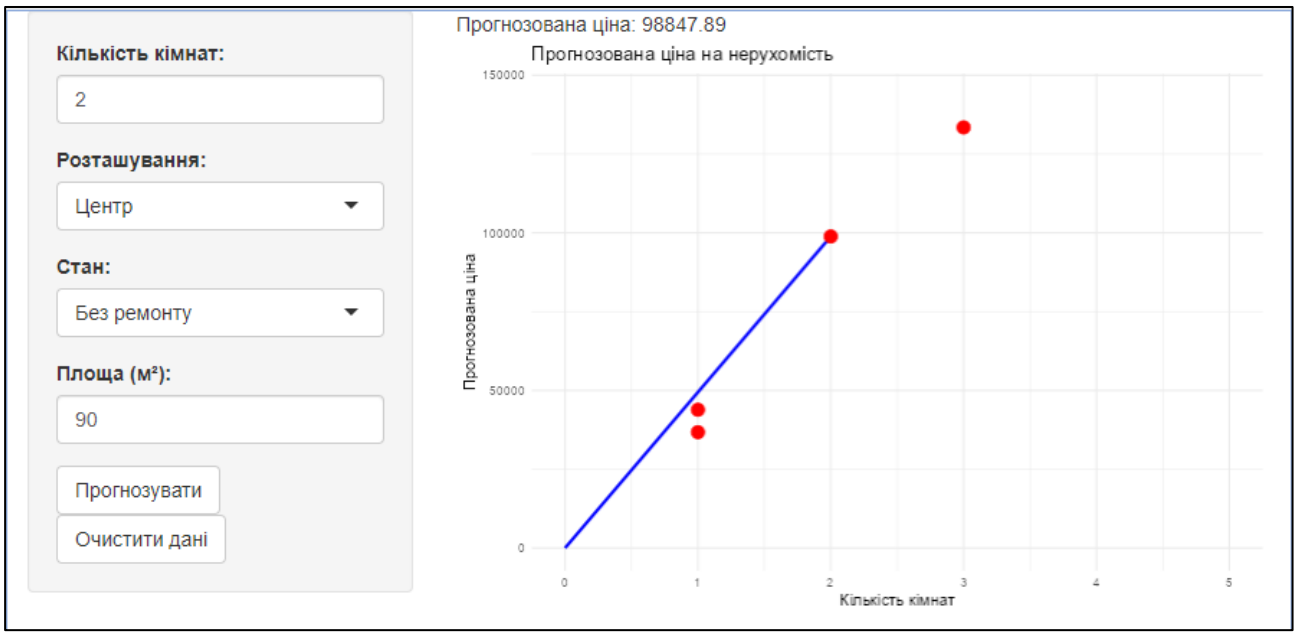

Рисунок 4.26 – Прогнозування ціни для об’єкта нерухомості

Графіки, що генеруються на основі введених даних, демонструють залежність прогнозованої ціни від кількості кімнат, відображаючи її на осі абсцис, а прогнозовану вартість – на осі ординат. Точки на графіку показують введені параметри, що допомагає користувачеві краще зрозуміти вплив кожного параметра на кінцеву вартість нерухомості.

Так, для нерухомості з більшою площею, навіть за інших незмінних параметрів, спостерігається підвищення прогнозованої ціни, що підтверджується розміщенням відповідних точок на графіку вище відносно осі ординат (див. рис. 4.27). Зміна стану з "Без ремонту" на "Ремонт" також підвищує вартість, відображаючи додану цінність об'єкта з кращими умовами.

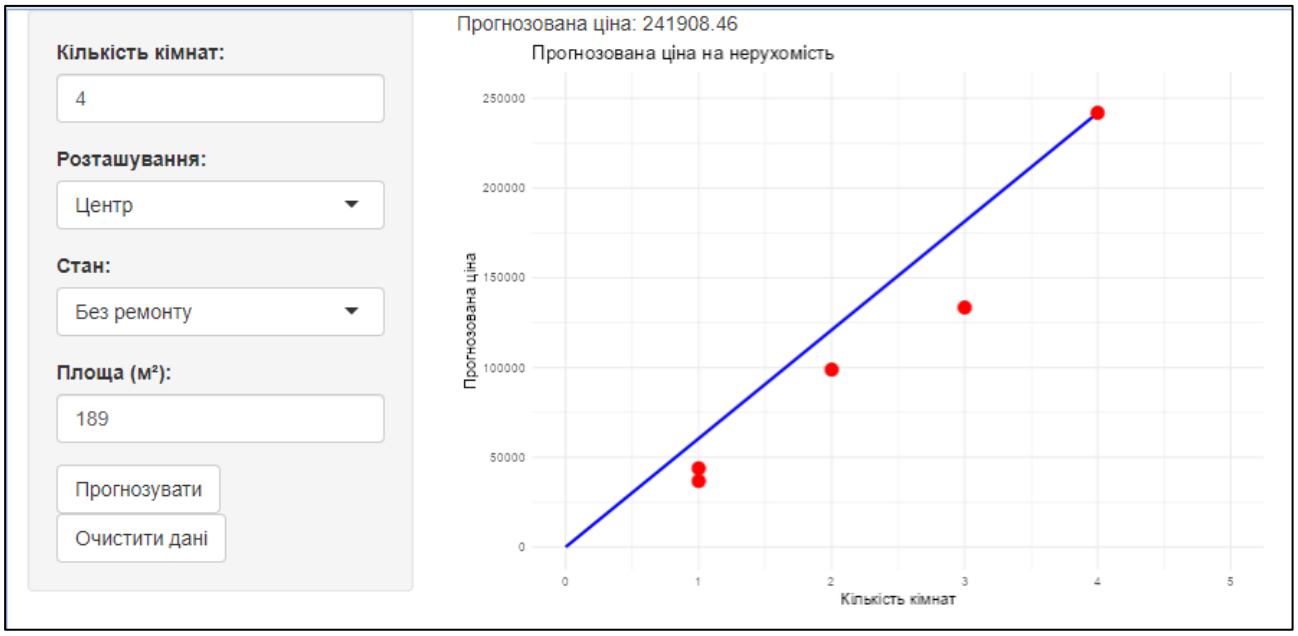

Рисунок 4.27 – Прогнозування ціни для об’єкта нерухомості

Таким чином, система не лише генерує точні прогнози, але й дозволяє користувачу аналізувати залежність ціни від різних факторів завдяки інтерактивному графічному представленню. Це значно підвищує її практичну цінність і зручність використання для оцінки вартості нерухомості в реальних ринкових умовах. Результати тестування системи прогнозування цін на ринку нерухомості демонструють її ефективність і відповідність поставленим цілям. Інтерфейс програми є інтуїтивно зрозумілим і зручним для користувачів, що дозволяє легко вводити параметри об’єкта та отримувати результати прогнозів. Система адекватно реагує на зміну вхідних даних, що підтверджується динамікою зміни прогнозованої ціни на графіках.

Графічне відображення результатів сприяє більш глибокому розумінню взаємозв’язку між параметрами нерухомості та її вартістю, що робить систему не лише засобом для прогнозування, але й інструментом для аналізу ринкових тенденцій. Завдяки реалізованій функціональності та перевіреним метрикам точності, система готова до використання в реальних умовах і здатна забезпечити користувачів обґрунтованими та надійними прогнозами.

Висновки до розділу 4

У четвертому розділі роботи реалізовано інформаційну систему для прогнозування цін на ринку нерухомості, яка базується на моделі множинної регресії. Після детальної оцінки якості кількох моделей, включаючи лінійну регресію, поліноміальну регресію, дерева рішень, випадковий ліс і XGBoost, модель множинної регресії була обрана як найбільш оптимальна завдяки її високій точності та простоті інтерпретації.

Система дозволяє враховувати ключові фактори, що впливають на формування цін, такі як площа, кількість кімнат, стан ремонту та розташування. Тестування показало, що прогнозовані ціни адекватно змінюються залежно від введених параметрів, що підтверджується графічною візуалізацією результатів. Це дозволяє користувачам аналізувати залежності між характеристиками нерухомості та її вартістю.

Розроблений інтерфейс забезпечує зручність введення даних і демонструє результат прогнозування у вигляді числових значень та графіків. Такий підхід робить систему доступною для широкого кола користувачів, включаючи осіб без технічної підготовки.

Таким чином, у четвертому розділі показано, що розроблена інформаційна система на основі моделі множинної регресії є надійним інструментом для прогнозування цін на ринку нерухомості. Вона забезпечує точність прогнозів і зручність використання, що робить її готовою до практичного впровадження.

ВИСНОВКИ

У кваліфікаційній роботі було розроблено інформаційну систему прогнозування цін на ринку нерухомості, яка базується на сучасних методах аналізу даних і машинного навчання. Робота складалася з кількох етапів, кожен з яких забезпечував досягнення основної мети – створення точного, надійного та зручного інструменту для оцінки ринкової вартості нерухомості.

У першому розділі проведено аналіз предметної області, визначено основні чинники, що впливають на ціноутворення нерухомості, такі як площа, стан ремонту, кількість кімнат і розташування. Також було виконано огляд існуючих інформаційних систем і методів прогнозування цін. Встановлено, що впровадження сучасних підходів, таких як машинне навчання, значно підвищує точність прогнозів і знижує ризики для інвесторів та інших учасників ринку.

Другий розділ був присвячений побудові структури інформаційної системи та вибору методів для вирішення задачі. Детально описано методи обробки даних, нормалізації, кодування категоріальних змінних, а також алгоритми оцінки якості моделей. Структура системи розроблена таким чином, щоб забезпечити її функціональність, модульність та можливість адаптації до зміни ринкових умов.

У третьому розділі виконано аналіз та попередню обробку даних. Здійснено очистку даних від пропущених значень, аномалій, а також нормалізацію та стандартизацію змінних. Також реалізовано кодування категоріальних змінних для подальшого їх використання в моделі. Проведено аналіз кореляцій між змінними для визначення найбільш значущих факторів, що впливають на формування ціни.

У четвертому розділі реалізовано інформаційну систему, протестовано її функціональність і оцінено точність прогнозів. На основі порівняння кількох моделей, таких як лінійна регресія, поліноміальна регресія, множинна регресія, дерева рішень, випадковий ліс та XGBoost, було обрано модель множинної регресії. Тестування системи підтвердило її точність і здатність адекватно реагувати на зміни вхідних параметрів. Інтерфейс системи забезпечує зручність введення даних

і візуалізацію результатів, що робить її доступною навіть для користувачів без спеціалізованих технічних знань.

У підсумку, виконана робота продемонструвала, що розроблена інформаційна система відповідає поставленим цілям та є ефективним інструментом для прогнозування цін на ринку нерухомості. Система не лише забезпечує високу точність прогнозів, але й дозволяє аналізувати залежності між характеристиками нерухомості та її вартістю, що може бути корисним як для індивідуальних користувачів, так і для бізнесу.

Результати роботи можуть бути використані для прийняття рішень щодо купівлі, продажу чи інвестування в об'єкти нерухомості. Також вони створюють основу для подальших досліджень у напрямку використання більш складних моделей та методів машинного навчання, а також інтеграції системи з реальними базами даних ринку нерухомості.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Ключка О. В., Лісовська А. С. Класифікація та структура сучасного ринку нерухомості як складової фінансового ринку // Економіка та суспільство. – 2023. – Вип. 57. – С. 457-458. URL: https://doi.org/10.32782/2524-0072/2023-57-119

2. Кучеренко В. Р., Заєць М. А., Захарченко О. В., Сментина Н. В., Улибіна В. О. Оцінка та управління нерухомістю: навчальний посібник. – Одеса: Видавництво ТОВ «Лерадрук». – 272 с.

3. Тенденції ринку нерухомості України. YouControl. URL: https://blog.youcontrol.market/tiendientsiyi-rinku-nierukhomosti-ukrayini/ (дата звернення: 10.08.2024).

4. Ціни на квартири в Києві. DOM.RIA. URL: https://dom.ria.com/uk/prodazha-kvartir/kiev/ceny/ (дата звернення: 20.10.2024).

5. Ціни на квартири в Одесі. DOM.RIA. URL: https://dom.ria.com/uk/prodazha-kvartir/odessa/ceny/ (дата звернення: 15.08.2024).

6. Ціни на квартири у Львові. DOM.RIA. URL: https://dom.ria.com/uk/prodazha-kvartir/lvov/ceny/ (дата звернення: 20.08.2024).

7. Державна служба статистики України. Статистичний огляд соціально-економічного становища України. – Київ, 2021.

8. Федоренко О. С. Стан ринку нерухомості в Україні: аналітика LUN // НВ Бізнес. URL: https://biz.nv.ua/ukr/consmarket/stan-rinku-neruhomosti-v-ukrajini-analitika-lun-50403512.html (дата звернення: 01.09.2024).

9. Бурматов О. О., Чугунов А. Г. A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments // Advances in Electrical and Electronic Engineering. 2023.

10. Mora-Garcia, R.-T., Cespedes-Lopez, M.-F., & Perez-Sanchez, V. R. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. *Land*, 11(21), 2100. URL: https://doi.org/10.3390/land11112100

11. Bhandari M., Mishra B., Ghosh S. Estimating Stock Market Prices with Histogram-based Gradient Boosting Regressor: A Case Study on Alphabet Inc. // Journal of Computational Finance. 2022.

12. Li C. Housing price prediction using machine learning // School of International Education, GuangDong University of Technology. 2024. URL: https://doi.org/10.54254/2755-2721/53/20241426 (дата звернення: 10.09.2024).

13. Breiman, L. (2021). Random Forests. Machine Learning, 45(1), 5-32. URL: 10.1023/A:1010933404324.

14. Zhang, Y., & Singer, B. (2021). Random Forest: A Classification and Regression Tool for Clinical Research. Statistical Medicine, 29(5), 498-511. URL: 10.1002/sim.3949.

15. Khoshgoftaar T. M., Gao M. House price prediction using machine learning // Expert Systems with Applications. 2020.

16. Hougen, C. D., Kaplan, L. M., Cerutti, F., & Hero, A. O. III. (2022). Uncertain Bayesian Networks: Learning from Incomplete Data. arXiv preprint arXiv:2208.04221. URL: https://arxiv.org/abs/2208.04221.

17. Dadalt, E., Romanelli, M., Pichler, G., & Piantanida, P. (2023). A Data-Driven Measure of Relative Uncertainty for Misclassification Detection. arXiv preprint arXiv:2306.01710. URL: https://arxiv.org/abs/2306.01710.

18. Smith, J., & Brown, A. (2022). Standardization and Normalization of Data: A Comparative Study. Journal of Data Science and Analytics, 15(3), 45-60.

19. El Morr, C., Jammal, M., Ali-Hassan, H., & El-Hallak, W. (2022). Data Preprocessing. In Machine Learning for Practical Decision Making (pp. 117–163). Springer. URL: https://link.springer.com/chapter/10.1007/978-3-031-41933-1_6.

20. Lozano-Murcia, C., Romero, F. P., Serrano-Guerrero, J., & Olivas, J. A. (2023). A comparison between explainable machine learning methods for classification and regression problems in the actuarial context. *Mathematics*, 11(14), 3088.

21. James G, Witten D, Hastie T, Tibshirani R, Taylor J (2023) Linear regression. An introduction to statistical learning: with applications in python. Springer International Publishing, Cham, pp 69–134.

22. Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: empirical evidence using real-world microdata. *Machine Learning and Applications*, 5, 100074.

23. Perkins, S., Davis, H., & Preez, V. D. (2020). Practical data science for actuarial tasks. A practical example of data science considerations by Modelling, Analytics and Insights in Data working party–New approaches to current actuarial work. *Journal of Actuarial Practice*.

24. Bai, J., Li, Y., Li, J., Yang, X., Jiang, Y., & Xia, S. T. (2022). Multinomial random forest. Pattern Recognition, 122, 108331.URL: https://www.sciencedirect.com/science/article/abs/pii/S0031320321005112?via%3Dihub

25. Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., & Geng, Y. A. (2022). Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, 276.

26. Liu, X., Li, Y., & Jiang, J. (2021). Simple measures of uncertainty for model selection. *TEST*, 30, 673–692. URL: https://doi.org/10.1007/s11749-020-00737-9

27. Data Preprocessing Techniques in Machine Learning: Steps & Best Practices // Towards Data Science. 2022.

28. Shahrivari A. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data // Journal of Building Performance. 2021.

29. Data Preprocessing Techniques: 6 Steps to Clean Data in Machine Learning // Analytics Vidhya. 2022.

30. Kumar A. Data Preprocessing: Steps, Techniques, and Importance in Machine Learning // International Journal of Computer Applications. 2021.2

31. Bhat, G. R. N. K., Kumar, H., & Patil, A. M. A Comprehensive Guide to Data Preprocessing // Research Journal of Applied Sciences, Engineering and Technology. – 2021. – URL: 10.19026/rjaset.12.100006.

32. Train in Data Team. Mastering Data Preprocessing: Techniques and Best Practices. – URL: https://www.blog.trainindata.com/mastering-data-preprocessing-techniques-and-best-practices (дата звернення: 20.09.2024).

33. Nedashkovskaya N.I., Lupanenko S.O. Comparative analysis of machine learning models for forecasting COVID-19 spreading in different countries // Èlektronne modeljuvannja. – 2020. – Vol. 42, № 5. – P. 51–65. – URL: https://doi.org/10.15407/emodel.42.05.051.

34. Chakraborty, T., Ghosh, I. (2020), “Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis”, *Chaos, Solitons & Fractals*, Vol. 135. DOI: 10.1016/j.chaos.2020.109850.

35. Ordóñez C., Lasheras F.S., Roca-Pardiñas J., de Cos Juez F.J. A hybrid ARIMA–SVM model for the study of the remaining useful life of aircraft engines // Journal of Computational and Applied Mathematics. – 2019. – Vol. 346. – P. 184–191. – DOI: https://doi.org/10.1016/j.cam.2018.06.032.

36. Petropoulos F., Makridakis S. Forecasting the novel coronavirus COVID-19 // PLoS ONE. – 2020. – Vol. 15, № 3. – P. e0231236. – DOI: https://doi.org/10.1371/journal.pone.0231236.

37. Abumohsen M., Owda A.Y., Owda M. Electrical Load Forecasting Based on Random Forest, XGBoost, and Linear Regression // IEEE Xplore. – 2023. – URL: https://ieeexplore.ieee.org/document/10225968.

38. Sorrentino N., Menniti D., Pinnarelli A., Brusco G., Vizza P., Mendicino S. A Combined ML Forecast and Real Time Control Approach to Reduce Imbalance Costs in Renewable Energy Communities of Prosumagers // 2024 AEIT International Annual Conference (AEIT). – 2024.

39. Ibrahem M.I., Fouda M.M. A Lightweight Privacy-Preserving Load Forecasting and Monitoring Scheme Supporting Dynamic Billing for Smart Grids: No KDC Required // IEEE Internet of Things Journal. – 2024. – Vol. 11, № 19.

40. Oshodi I. Machine Learning-Based Algorithms for Weather Forecasting. *International Journal of Artificial Intelligence and Machine Learning*. 2022. Vol. 2, no. 2. P. 12–20. URL: https://doi.org/10.51483/ijaiml.2.2.2022.12-20.

41. Castle J. L., Doornik J. A., Hendry D. F. Forecasting Principles from Experience with Forecasting Competitions. *Forecasting*. 2021. Vol. 3, no. 1. Ст. 138–165. URL: https://doi.org/10.3390/forecast3010010.

42. A Comparison of Flare Forecasting Methods. IV. Evaluating Consecutive-day Forecasting Patterns / S. H. Park et al. 2019. URL: http://hdl.handle.net/10454/18411 .

43. Forecasting International Stock Market Trends: XGBoost, LSTM, LSTM-XGBoost, and Backtesting XGBoost Models / H. OUKHOUYA et al. *Statistics, Optimization & Information Computing*. 2023. Vol. 12, no. 1. Ст. 200–209. URL: https://doi.org/10.19139/soic-2310-5070-1822 .

44. Generalized Linear Models / L. Fahrmeir та ін. *Regression*. Berlin, Heidelberg, 2021. С. 283–342. URL: https://doi.org/10.1007/978-3-662-63882-8_5.

45. Kahane L. H. The Multiple Regression Model. *Regression Basics*. London, 2024. С. 63–87. URL: https://doi.org/10.4324/9781003349174-5.

ДОДАТОК А

Код для попередньої обробки даних

```
f<-read.csv2('flats.csv',header = TRUE, encoding = 'UNICOD')
library(psych)
describe(f)
f <- f[f$price < 300000, ]
f <- f[f$price > 10000, ]
describe(f[, c('rooms', 'm2', 'price')])
library(ggplot2)
par(mfrow = c(1, 3))
hist(f$rooms, col = 'dark blue', main = 'rooms', xlab = 'Value')
hist(f$m2, col = 'dark blue', main = 'm2', xlab = 'Value')
hist(f$price, col = 'dark blue', main = 'price', xlab = 'Value')
any(is.na(f))
colSums(is.na(f[, c("rooms", "type")]))
f_clean <- f[!is.na(f$rooms) & !is.na(f$type), ]
any(is.na(f_clean))
par(mfrow = c(1, 3))
boxplot(f_clean$rooms)
boxplot(f_clean$m2)
boxplot(f_clean$price)
qplot(data =f_clean, x=condition, y=price, geom = "boxplot" )
qplot(data =f_clean, x=location, y=price, geom = "violin" )

## Preprocessing
# Factors as numeric
f_clean$location <- as.numeric(as.factor(f_clean$location)) - 1
f_clean$condition <- as.numeric(as.factor(f_clean$condition)) - 1
f_clean$type <- as.numeric(as.factor(f_clean$type)) - 1

# Missing data
f_clean$rooms <- ifelse(is.na(f_clean$rooms),
                round(mean(f_clean$rooms, na.rm = TRUE)), f_clean$rooms)
f_clean$type <- ifelse(is.na(f_clean$type),
               round(mean(f_clean$type, na.rm = TRUE)), f_clean$type)
```

```
## Visualising
library(ggplot2)
par(mfrow = c(2, 3))

hist(f_clean$rooms, col = 'dark blue', main = 'rooms', xlab = 'Value')
hist(f_clean$m2, col = 'dark blue', main = 'm2', xlab = 'Value')
hist(f_clean$price, col = 'dark blue', main = 'price', xlab = 'Value')

hist(log(f_clean$rooms), col = 'dark blue', main = 'rooms', xlab = 'Value')
hist(log(f_clean$m2), col = 'dark blue', main = 'm2', xlab = 'Value')
hist(log(f_clean$price), col = 'dark blue', main = 'price', xlab = 'Value')

## log
f_clean$rooms <- log(f_clean$rooms)
f_clean$m2 <- log(f_clean$m2)
f_clean$price <- log(f_clean$price)
describe(f_clean[, c('rooms', 'm2', 'price')])

# Replace ejections with max (no need)
f_clean$rooms <- ifelse(f_clean$rooms < mean(f_clean$rooms) + sd(f_clean$rooms) *
3,
                        f_clean$rooms, mean(f_clean$rooms) + sd(f_clean$rooms) * 3)
f_clean$rooms <- ifelse(f_clean$rooms > mean(f_clean$rooms) - sd(f_clean$rooms) *
3,
                        f_clean$rooms, mean(f_clean$rooms) - sd(f_clean$rooms) * 3)

f_clean$price <- ifelse(f_clean$price < mean(f_clean$price) + sd(f_clean$price) * 3,
                        f_clean$price, mean(f_clean$price) + sd(f_clean$price) * 3)
f_clean$price <- ifelse(f_clean$price > mean(f_clean$price) - sd(f_clean$price) * 3,
                        f_clean$price, mean(f_clean$price) - sd(f_clean$price) * 3)

f_clean$m2 <- ifelse(f_clean$m2 < mean(f_clean$m2) + sd(f_clean$m2) * 3,
                     f_clean$m2, mean(f_clean$m2) + sd(f_clean$m2) * 3)
f_clean$m2 <- ifelse(f_clean$m2 > mean(f_clean$m2) - sd(f_clean$m2) * 3,
                     f_clean$m2, mean(f_clean$m2) - sd(f_clean$m2) * 3)

describe(f_clean[, c('rooms', 'm2', 'price')])
```

```
# Завантажуємо необхідні бібліотеки
library(GGally)
library(ggplot2)

# Створюємо розширену матрицю розсіювання для змінних rooms, m2, price
ggpairs(f_clean[, c('rooms', 'm2', 'price')],
        upper = list(continuous = wrap("cor", size = 6)),  # Кореляції у верхній частині
        lower = list(continuous = "smooth"),  # Діаграми розсіювання з лініями тренду
        diag = list(continuous = "barDiag"))  # Гістограми на діагоналі
```

ДОДАТОК Б

Код методу лінійної регресії

```
## Download the data
#Download the files
f_train <- read.csv2('flats_train.csv', header = TRUE, encoding = 'UNICOD')
f_train <- f_train[, -1]
f_test <- read.csv2('flats_test.csv', header = TRUE, encoding = 'UNICOD')
f_test <- f_test[, -1]
# Correlations
library(psych)
pairs.panels(f_train, lm=TRUE, # linear fit
             method = "pearson", # correlation method
             hist.col = "#00AFBB")

## Simple Linear Regression (one factor - m2)
# Fitting Simple Linear Regression to the Training set
sr <- lm(price ~ m2, f_train)
summary(sr)

## Predicting
p_sr <- predict(sr, f_test)

# MSE
train_mse_sr <- sum((f_train$price - predict(sr, f_train))^2) / length(f_train$price)
test_mse_sr <- sum((f_test$price - p_sr)^2) / length(p_sr)

train_mse_sr
test_mse_sr

## Visualising
library(ggplot2)

ggplot() +
  geom_point(aes(f_train$m2, f_train$price), colour = 'red') +
  geom_point(aes(f_test$m2, f_test$price), colour = 'dark green') +
  geom_line(aes(f_test$m2, p_sr), colour = 'blue') +
  ggtitle('Price vs m2') +
  xlab('m2') +
  ylab('price')
```

ДОДАТОК В

Код методу множинної лінійної регресії

```
## Multiple Linear Regression (many factors)
# All factors
mr <- lm(price ~ ., f_train)
summary(mr)

## Optimized model
# as p-value, Pr(>|t|) of variable "type" is higher than significance level (5%), let's exclude this variable from the model
mr_opt <- lm(price ~ rooms + location + condition + m2, f_train)
summary(mr_opt)

# Prediction
p_mr <- predict(mr_opt, f_test)
train_mse_opt <- sum((f_train$price - predict(mr_opt, f_train))^2) / length(f_train$price)
test_mse_opt <- sum((f_test$price - p_mr)^2) / length(p_mr)

train_mse_opt
test_mse_opt

# Visualising
ggplot() +
  geom_point(aes(f_train$m2, f_train$price), colour = 'red') +
  geom_point(aes(f_test$m2, f_test$price), colour = 'dark green') +
  geom_line(aes(f_test$m2, p_mr), colour = 'blue') +
  ggtitle('Price vs m2') +
  xlab('m2') +
  ylab('price')
```

ДОДАТОК Г

Код методу поліноміальної регресії

```
## Polynomial Linear Regression (one factor - m2)
# Features extending
f_train_poly <- f_train[, c('price', 'm2')]
f_test_poly <- f_test[, c('price', 'm2')]

f_train_poly$m2_squared <- f_train_poly$m2^2
f_train_poly$m2_cubed <- f_train_poly$m2^3

f_test_poly$m2_squared <- f_test_poly$m2^2
f_test_poly$m2_cubed <- f_test_poly$m2^3

## 3 powers
pr <- lm(price ~ m2 + m2_squared + m2_cubed, f_train_poly)
summary(pr)


# Predicting
p_pr <- predict(pr, f_test_poly)
train_mse_poly <- sum((f_train_poly$price - predict(pr, f_train_poly))^2) /
length(f_train_poly$price)
test_mse_poly <- sum((f_test_poly$price - p_pr)^2) / length(p_pr)

train_mse_poly
test_mse_poly

# Visualising
ggplot() +
  geom_point(aes(f_train_poly$m2, f_train_poly$price), colour = 'red') +
  geom_point(aes(f_test_poly$m2, f_test_poly$price), colour = 'dark green') +
  geom_line(aes(f_test_poly$m2, p_pr), colour = 'blue') +
  ggtitle('Price vs m2') +
  xlab('m2') +
  ylab('price')
```

ДОДАТОК Д

Код методу дерев рішень

```
# Download the data
# Download the files
f_train <- read.csv2('flats_train.csv', header = TRUE, encoding = 'UNICOD')
f_test <- read.csv2('flats_test.csv', header = TRUE, encoding = 'UNICOD')

# Decision Tree Regression
# Fitting simple tree
install.packages('rpart')
library(rpart)

dt <- rpart(price ~ m2, f_train, control = rpart.control(minsplit = 50))
plot(dt)
text(dt, pos = 1, cex = .75, col = 1, font = 1)

# Predicting
p_dt <- predict(dt, f_test)
train_mse_dt <- sum((f_train$price - predict(dt, f_train))^2) / length(f_train$price)
test_mse_dt <- sum((f_test$price - p_dt)^2) / length(p_dt)

train_mse_dt
test_mse_dt

# Visualising
library(ggplot2)

x_grid <- seq(min(f_train$m2), max(f_train$m2), 0.01)
ggplot() +
  geom_point(aes(f_train$m2, f_train$price), colour = 'red') +
  geom_point(aes(f_test$m2, f_test$price), colour = 'dark green') +
  geom_line(aes(x_grid, predict(dt, data.frame(m2 = x_grid))), colour = 'blue') +
  ggtitle('Price vs m2') +
  xlab('m2') +
  ylab('price')
```

ДОДАТОК Е

Код методу випадкового лісу

```
## Random forest
# Fitting
install.packages('randomForest')
library(randomForest)

set.seed(1234)
rf <- randomForest(x = f_train['m2'], y = f_train$price, ntree = 5)

# Predicting
p_rf <- predict(rf, f_test)
train_mse_rf <- sum((f_train$price - predict(rf, f_train))^2) / length(f_train$price)
test_mse_rf <- sum((f_test$price - p_rf)^2) / length(p_rf)

train_mse_rf
test_mse_rf

# Visualising
ggplot() +
  geom_point(aes(f_train$m2, f_train$price), colour = 'red') +
  geom_point(aes(f_test$m2, f_test$price), colour = 'dark green') +
  geom_line(aes(x_grid, predict(rf, data.frame(m2 = x_grid))), colour = 'blue') +
  ggtitle('Price vs m2') +
  xlab('m2') +
  ylab('price')

# Saving results
fit_1 <- read.csv2('flats_fit_1.csv', header = TRUE)
fit_1$p_dt <- p_dt
fit_1$p_rf <- p_rf
write.table(fit_1, file = 'flats_fit_1.csv', sep = ";", row.names = FALSE, col.names =
TRUE, fileEncoding = 'UTF-8')
head(fit_1)
```

ДОДАТОК Ж

Код методу XBoost

```
install.packages("xgboost")
library(xgboost)
library(ggplot2)

f_train <- read.csv2('flats_train.csv', header = TRUE)
f_test <- read.csv2('flats_test.csv', header = TRUE)

dtrain <- xgb.DMatrix(data = as.matrix(f_train[, c('m2', 'rooms', 'location', 'condition')]),
label = f_train$price)
dtest <- xgb.DMatrix(data = as.matrix(f_test[, c('m2', 'rooms', 'location', 'condition')]),
label = f_test$price)

params <- list(
  objective = "reg:squarederror", # Цільова функція для регресії
  eta = 0.1,                    # Швидкість навчання
  max_depth = 5,                # Максимальна глибина дерева
  eval_metric = "rmse"          # Метрика для оцінки моделі
)

set.seed(1234) # Для відтворюваності результатів
nrounds <- 100 # Кількість раундів навчання
xgb_model <- xgb.train(params, dtrain, nrounds)

p_xgb <- predict(xgb_model, dtest)

# Обчислення середньої квадратичної помилки (MSE)
train_mse_xgb <- sum((f_train$price - predict(xgb_model, dtrain))^2) /
length(f_train$price)
test_mse_xgb <- sum((f_test$price - p_xgb)^2) / length(p_xgb)

# Виведення результатів MSE
train_mse_xgb
test_mse_xgb

# Візуалізація результатів
ggplot() +
  geom_point(aes(f_train$m2, f_train$price), colour = 'red') +
  geom_point(aes(f_test$m2, f_test$price), colour = 'dark green') +
```

```
  geom_line(aes(x = f_test$m2, y = p_xgb), colour = 'blue') +
  ggtitle('Price vs m2 (XGBoost)') +
  xlab('m2') +
  ylab('price')

fit_1 <- read.csv2('flats_fit_1.csv', header = TRUE)
fit_1$p_xgb <- p_xgb
write.table(fit_1, file = 'flats_fit_1.csv', sep = ";", row.names = FALSE, col.names =
TRUE, fileEncoding = 'UTF-8')
head(fit_1)
```

ДОДАТОК И

Код порівняння якості моделей та прогнозів

```
write.csv2(fit_1[-1], file = "flats_fit_1.csv")

g_sr <- ggplot(fit_1, aes(x=f_test.price, y=p_sr)) +
  geom_abline(intercept=0, slope=1) +
  geom_point(alpha=0.5) +
  labs(title="Linear Regression", x="Real Price", y="Predicted Price") +
  theme(plot.title=element_text(size=10),
        axis.title.x=element_text(size=7),
        axis.title.y=element_text(size=7),
        axis.text.x=element_text(size=5),
        axis.text.y=element_text(size=5)) +
  theme(legend.position="none")

g_mr <- ggplot(fit_1, aes(x=f_test.price, y=p_mr)) +
  geom_abline(intercept=0, slope=1) +
  geom_point(alpha=0.5) +
  labs(title="Multiple Regression", x="Real Price", y="Predicted Price") +
  theme(plot.title=element_text(size=10),
        axis.title.x=element_text(size=7),
        axis.title.y=element_text(size=7),
        axis.text.x=element_text(size=5),
        axis.text.y=element_text(size=5)) +
  theme(legend.position="none")

g_pr <- ggplot(fit_1, aes(x=f_test.price, y=p_pr)) +
  geom_abline(intercept=0, slope=1) +
  geom_point(alpha=0.5) +
  labs(title="Polynomial Regression", x="Real Price", y="Predicted Price") +
  theme(plot.title=element_text(size=10),
        axis.title.x=element_text(size=7),
        axis.title.y=element_text(size=7),
        axis.text.x=element_text(size=5),
        axis.text.y=element_text(size=5)) +
  theme(legend.position="none")

g_dt <- ggplot(fit_1, aes(x=f_test.price, y=p_dt)) +
  geom_abline(intercept=0, slope=1) +
  geom_point(alpha=0.5) +
```

```
  labs(title="Regression Tree", x="Real Price", y="Predicted Price") +
  theme(plot.title=element_text(size=10),
        axis.title.x=element_text(size=7),
        axis.title.y=element_text(size=7),
        axis.text.x=element_text(size=5),
        axis.text.y=element_text(size=5)) +
  theme(legend.position="none")

g_rf <- ggplot(fit_1, aes(x=f_test.price, y=p_rf)) +
  geom_abline(intercept=0, slope=1) +
  geom_point(alpha=0.5) +
  labs(title="Random Forest", x="Real Price", y="Predicted Price") +
  theme(plot.title=element_text(size=10),
        axis.title.x=element_text(size=7),
        axis.title.y=element_text(size=7),
        axis.text.x=element_text(size=5),
        axis.text.y=element_text(size=5)) +
  theme(legend.position="none")

g_xgb <- ggplot(fit_1, aes(x=f_test.price, y=p_xgb)) +
  geom_abline(intercept=0, slope=1) +
  geom_point(alpha=0.5) +
  labs(title="XGBoost",x="Real Price",y="Predicted Price") +
  theme(plot.title=element_text(size=10),
        axis.title.x=element_text(size=7),
        axis.title.y=element_text(size=7),
        axis.text.x=element_text(size=5),
        axis.text.y=element_text(size=5)) +
  theme(legend.position="none")

gridExtra::grid.arrange(g_sr, g_mr, g_pr, g_dt, g_rf,g_xgb, ncol=2)

# Обчислення помилок прогнозування
sr <- mean((fit_1$f_test.price - fit_1$p_sr) ^ 2)
mr <- mean((fit_1$f_test.price - fit_1$p_mr) ^ 2)
pr <- mean((fit_1$f_test.price - fit_1$p_pr) ^ 2)
dt <- mean((fit_1$f_test.price - fit_1$p_dt) ^ 2)
rf <- mean((fit_1$f_test.price - fit_1$p_rf) ^ 2)
xgb <- mean((fit_1$f_test.price - fit_1$p_xgb) ^ 2)

mse <- data.frame(sr, mr, pr, dt, rf,xgb)
head(mse)
```

```
# Перетворення даних
mse1 <- melt(mse)
head(mse1)

# Побудова графіка
b1 <- ggplot(mse1, aes(x=variable, y=value)) +
  geom_bar(stat="summary", fun="mean", fill='royalblue') +
  labs(title="Середня квадратична похибка моделей",
       x="Моделі",
       y="Середня квадратична похибка") +
  theme_minimal() +
  theme(plot.title=element_text(size=12),
        axis.title.x=element_text(size=10),
        axis.title.y=element_text(size=10),
        axis.text.x=element_text(size=8, angle=45, hjust=1),
        axis.text.y=element_text(size=8))

# Відображення графіка
print(b1)

r_squared_results <- list()
models <- c("p_sr", "p_mr", "p_pr", "p_dt", "p_rf", "p_xgb")
# Обчислення R² для кожної моделі
for (model in models) {
  formula <- as.formula(paste("f_test.price ~", model))
  lm_model <- lm(formula, data = fit_1)  # Створення лінійної моделі
  r_squared_results[[model]] <- summary(lm_model)$r.squared  # Збереження R² в
список
}

r_squared_df <- data.frame(Model = names(r_squared_results), R_squared =
unlist(r_squared_results))
print(r_squared_df)

aic_results <- list()
models <- c("p_sr", "p_mr", "p_pr", "p_dt", "p_rf", "p_xgb")
# Обчислення AIC для кожної моделі
for (model in models) {
  formula <- as.formula(paste("f_test.price ~", model))
  lm_model <- lm(formula, data = fit_1)  # Створення лінійної моделі
  aic_results[[model]] <- AIC(lm_model)  # Збереження AIC в список
```

```
}

aic_df <- data.frame(Model = names(aic_results), AIC = unlist(aic_results))
print(aic_df)

install.packages("lmtest", dependencies=TRUE)
library(lmtest)
dw_results <- list()
models <- c("p_sr", "p_mr", "p_pr", "p_dt", "p_rf", "p_xgb")
# Обчислення DW для кожної моделі
for (model in models) {
  formula <- as.formula(paste("f_test.price ~", model))
  lm_model <- lm(formula, data = fit_1)
  dw_stat <- dwtest(lm_model)$statistic  # Обчислення статистики Дарбіна-Уотсона
  dw_results[[model]] <- dw_stat  # Збереження значення DW в список
}

dw_df <- data.frame(Model = names(dw_results), DW_statistic = unlist(dw_results))
print(dw_df)



fisher_results <- list()
models <- c("p_sr", "p_mr", "p_pr", "p_dt", "p_rf", "p_xgb")

# Обчислення F-статистики для кожної моделі
for (model in models) {
  formula <- as.formula(paste("f_test.price ~", model))
  lm_model <- lm(formula, data = fit_1)  #
  f_statistic <- summary(lm_model)$fstatistic[1]  # Обчислення F-статистики
  fisher_results[[model]] <- f_statistic  # Збереження значення F-статистики в список
}

fisher_df <- data.frame(Model = names(fisher_results), F_statistic =
unlist(fisher_results))
print(fisher_df)

# Функція для обчислення коефіцієнта Тейла
calculate_theil_u <- function(real, predicted) {
  numerator <- sqrt(mean((real - predicted)^2))
  denominator <- sqrt(mean(real^2)) + sqrt(mean(predicted^2))
  theil_u <- numerator / denominator
```

```
  return(theil_u)
}

# Створення списку для результатів Тейла
theil_results <- list()
models <- c("p_sr", "p_mr", "p_pr", "p_dt", "p_rf", "p_xgb")

# Обчислення коефіцієнта Тейла для кожної моделі
for (model in models) {
  theil_results[[model]] <- calculate_theil_u(fit_1$f_test.price, fit_1[[model]])
}

# Перетворення результатів у таблицю
theil_df <- data.frame(Model = names(theil_results), Theil_U = unlist(theil_results))

# Відображення результатів
print(theil_df)

# Побудова графіка для Тейла
theil_plot <- ggplot(theil_df, aes(x = Model, y = Theil_U)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Коефіцієнт Тейла для моделей",
       x = "Моделі",
       y = "Коефіцієнт Тейла") +
  theme_minimal() +
  theme(plot.title = element_text(size = 12),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(size = 8, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 8))

# Відображення графіка
print(theil_plot)

# Функція для обчислення MAPE
calculate_mape <- function(real, predicted) {
  mean(abs((real - predicted) / real)) * 100
}

# Створення списку для результатів MAPE
mape_results <- list()
models <- c("p_sr", "p_mr", "p_pr", "p_dt", "p_rf", "p_xgb")
```

```
# Обчислення MAPE для кожної моделі
for (model in models) {
  mape_results[[model]] <- calculate_mape(fit_1$f_test.price, fit_1[[model]])
}

# Перетворення результатів у таблицю
mape_df <- data.frame(Model = names(mape_results), MAPE = unlist(mape_results))

# Відображення результатів
print(mape_df)

# Побудова графіка для MAPE
mape_plot <- ggplot(mape_df, aes(x = Model, y = MAPE)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "MAPE для моделей",
       x = "Моделі",
       y = "MAPE (%)") +
  theme_minimal() +
  theme(plot.title = element_text(size = 12),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(size = 8, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 8))

# Відображення графіка
print(mape_plot)
```

ДОДАТОК К

Код для інтерфейсної частини системи

```
# Встановлення та завантаження необхідних пакетів
if (!require("shiny")) install.packages("shiny", dependencies = TRUE)
if (!require("ggplot2")) install.packages("ggplot2", dependencies = TRUE)

library(shiny)
library(ggplot2)

# Оптимізована модель (створена на логарифмованих даних)
mr_opt <- lm(price ~ rooms + location + condition + m2, f_train)

# Функція для прогнозування
predict_price_mr <- function(new_data) {
  predictions <- predict(mr_opt, newdata = new_data)
  return(predictions)
}

# UI частина
ui <- fluidPage(
  sidebarLayout(
    sidebarPanel(
      numericInput("rooms", "Кількість кімнат:", value = 1, min = 1),
      selectInput("location", "Розташування:",
                  choices = c("Віддалені райони" = 0, "Центр" = 1)),
      selectInput("condition", "Стан:",
                  choices = c("Без ремонту" = 0, "Ремонт" = 1)),
      numericInput("m2", "Площа (м²):", value = 50, min = 1),
      actionButton("predict", "Прогнозувати"),
      actionButton("clear", "Очистити дані")
    ),

    mainPanel(
      textOutput("prediction"),
      plotOutput("plot")
    )
  )
)

# Server частина
```

```
server <- function(input, output) {
  # Зберігаємо прогнози
  predictions <- reactiveVal(data.frame(rooms = numeric(0), prices = numeric(0)))

  observeEvent(input$predict, {
    # Логарифмування вхідних даних
    new_data <- data.frame(
      rooms = log(input$rooms),  # Логарифмуємо кількість кімнат
      location = as.integer(input$location),  # Категоріальні змінні
      condition = as.integer(input$condition),
      m2 = log(input$m2)  # Логарифмуємо площу
    )

    # Прогнозування (логарифмована ціна)
    predicted_price_log <- predict_price_mr(new_data)

    # Відновлення ціни до оригінального масштабу
    predicted_price <- exp(predicted_price_log)

    # Вивід прогнозованої ціни
    output$prediction <- renderText({
      paste("Прогнозована ціна:", round(predicted_price, 2))
    })

    # Додаємо новий прогноз до вектору
    current_predictions <- predictions()
    new_predictions <- rbind(current_predictions, data.frame(rooms = input$rooms,
prices = predicted_price))
    predictions(new_predictions)

    # Побудова графіка
    output$plot <- renderPlot({
      ggplot() +
        geom_segment(aes(x = 0, y = 0, xend = input$rooms, yend = predicted_price),
                     color = "blue", size = 1) +
        geom_point(data = predictions(), aes(x = rooms, y = prices), size = 4, color =
"red") +
        geom_segment(data = predictions(), aes(x = lag(rooms), y = lag(prices), xend =
rooms, yend = prices),
                     color = "blue", size = 1) +
        labs(title = "Прогнозована ціна на нерухомість",
             x = "Кількість кімнат",
```

```
          y = "Прогнозована ціна") +
      theme_minimal() +
      xlim(0, 5) +
      ylim(0, max(predictions()$prices) + 10000)
  })
 })

 # Обробка кнопки "Очистити дані"
 observeEvent(input$clear, {
  predictions(data.frame(rooms = numeric(0), prices = numeric(0)))  # Очищуємо
прогнози
  output$prediction <- renderText({
   "Прогнозована ціна:"
  })
  output$plot <- renderPlot({
   ggplot() +
    labs(
     x = "Кількість кімнат",
     y = "Прогнозована ціна") +
    theme_minimal()  # Порожній графік
  })
 })
}
# Запуск Shiny App
shinyApp(ui = ui, server = server)
```