

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук
(повна назва)
Кафедра _____ Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____ перший (бакалаврський)

Розробка інтелектуальної системи прогнозування вартості об'єктів
на основі методів машинного навчання
(тема)

Виконав:
здобувач _____ четвертого _____ року навчання,
групи _____ ІТШ-22-1

Анастасія ЛІЩЕНКО
(власне ім'я, прізвище)

Спеціальність 122 Комп'ютерні науки

(код і повна назва спеціальності)

Тип програми _____ освітньо-професійна

Освітня програма _____ Штучний інтелект

(повна назва освітньої програми)

Керівник _____ доц. Олег ЗОЛОТУХІН
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ

(підпис)

Лариса ЧАЛА
(власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____

Кафедра _____ Штучного інтелекту _____

Рівень вищої освіти _____ перший (бакалаврський) _____

Спеціальність _____ 122 Комп'ютерні науки _____
(код і повна назва)

Тип програми _____ освітньо-професійна _____

Освітня програма _____ Штучний інтелект _____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

« _____ » _____ 20 ____ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві _____ Ліщенко Анастасії Ігорівні _____
(прізвище, ім'я, по батькові)

1. Тема роботи _____ Розробка інтелектуальної системи прогнозування вартості об'єктів на основі методів машинного навчання _____

затверджена наказом університету від 5 червня 2026 р. № 488Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 10 липня 2026 р.

3. Вихідні дані до роботи _____ Наукові публікації з тематики машинного навчання, регресійного аналізу та прогнозування вартості нерухомості; документація бібліотек Python (scikit-learn, XGBoost, CatBoost, SHAP, NumPy, Pandas, Matplotlib); набір даних Ames Housing Dataset для проведення експериментального дослідження, навчання, тестування та порівняльного аналізу моделей машинного навчання _____

4. Перелік питань, що потрібно опрацювати в роботі _____

1) Аналіз предметної галузі та постановка задачі дослідження

2) Методи машинного навчання для прогнозування вартості об'єктів

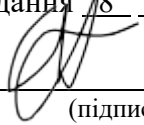
3) Розробка інтелектуальної системи прогнозування вартості об'єктів

4) Експериментальне дослідження інтелектуальної системи

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	08.06.2026	виконано
2	Аналіз предметної галузі та постановка задачі	08.06.2026–10.06.2026	виконано
3	Теоретичне дослідження методів машинного навчання для прогнозування вартості нерухомості	10.06.2026–12.06.2026	виконано
4	Формування експериментального набору даних та попередня обробка	13.06.2026–16.06.2026	виконано
5	Розробка ML pipeline та реалізація моделей машинного навчання	17.06.2026–23.06.2026	виконано
6	Проведення експериментального дослідження, оптимізація гіперпараметрів та інтерпретація результатів	24.06.2026–27.06.2026	виконано
7	Аналіз результатів експериментального дослідження	27.06.2026–29.06.2026	виконано
8	Оформлення пояснювальної записки	30.06.2026–03.07.2026	виконано
9	Нормоконтроль	04.07.2026–06.07.2026	виконано
10	Рецензування	06.07.2026	виконано
11	Попередній захист	08.07.2026	виконано
12	Захист перед ЕК	10.07.2026	

Дата видачі завдання 18 червня 2026 р.

Здобувач 
(підпис)

Керівник роботи _____
(підпис)

доц. Олег ЗОЛОТУХІН
(посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 93 с., 50 рис., 1 дод., 25 джерел.

ВИПАДКОВИЙ ЛІС, ГРАДІЄНТНИЙ БУСТИНГ, ІНТЕРПРЕТАЦІЯ МОДЕЛЕЙ, МАШИННЕ НАВЧАННЯ, ПРОГНОЗУВАННЯ ВАРТОСТІ НЕРУХОМОСТІ, РЕГРЕСІЯ, SHAP-АНАЛІЗ.

Об'єкт дослідження – процес прогнозування вартості об'єктів нерухомості на основі їх характеристик.

Предмет дослідження – методи машинного навчання для прогнозування вартості об'єктів нерухомості, а також підходи до інтерпретації отриманих результатів.

Мета роботи – розробка інтелектуальної системи прогнозування вартості об'єктів нерухомості на основі методів машинного навчання, дослідження ефективності моделей прогнозування та інтерпретація отриманих результатів.

Методи дослідження – методи попередньої обробки даних, формування ознак, методи регресійного аналізу, методи для оптимізації гіперпараметрів, статистичні метрики оцінювання якості моделей (MAE, RMSE, R^2), метод SHAP для інтерпретації моделей машинного навчання.

У роботі розроблено експериментальний ML pipeline прогнозування вартості об'єктів нерухомості, який охоплює етапи підготовки даних, формування додаткових ознак, навчання моделей машинного навчання, їх порівняльного оцінювання, оптимізації гіперпараметрів та інтерпретації отриманих прогнозів. За результатами дослідження визначено найбільш ефективну модель прогнозування та виконано аналіз факторів, що найбільше впливають на формування прогнозованої вартості об'єктів нерухомості із застосуванням SHAP-аналізу.

ABSTRACT

Bachelor's thesis contains: 93 pp., 50 fig., 1 ann., 25 references.

GRADIENT BOOSTING, MACHINE LEARNING, MODEL INTERPRETABILITY, RANDOM FOREST, REAL ESTATE PRICE PREDICTION, REGRESSION, SHAP ANALYSIS.

Object of research – the process of predicting real estate prices based on property characteristics.

Subject of research – machine learning methods for real estate price prediction and approaches to the interpretation of the obtained results.

Purpose of the study – to develop an intelligent real estate price prediction system based on machine learning methods, investigate the effectiveness of prediction models, and interpret the obtained results.

Research methods – data preprocessing methods, feature engineering methods, regression analysis methods, hyperparameter optimization methods, statistical model evaluation metrics (MAE, RMSE, R^2), and the SHAP method for machine learning model interpretation.

The study presents an experimental ML pipeline for real estate price prediction, including data preprocessing, feature engineering, machine learning model training, comparative model evaluation, hyperparameter optimization, and interpretation of prediction results. The study identifies the most effective prediction model and analyzes the factors that have the greatest influence on the predicted real estate prices using SHAP analysis.

ЗМІСТ

Вступ.....	7
1 Аналіз предметної галузі та постановка задачі дослідження	8
1.1 Особливості формування вартості об'єктів нерухомості	8
1.2 Аналіз існуючих підходів і систем прогнозування вартості	13
1.3 Постановка задачі дослідження.....	30
2 Методи машинного навчання для прогнозування вартості об'єктів.....	32
2.1 Регресійні методи машинного навчання	32
2.2 Ансамблеві методи машинного навчання	36
2.3 Метрики оцінювання якості регресійних моделей.....	43
2.4 Методи інтерпретації результатів машинного навчання	46
3 Розробка інтелектуальної системи прогнозування вартості об'єктів	49
3.1 Архітектура інтелектуальної системи прогнозування	50
3.2 Формування експериментального набору даних.....	52
3.3 Попередня обробка даних та формування ознак	55
3.4 Реалізація моделей машинного навчання та експериментального дослідження	60
4 Експериментальне дослідження інтелектуальної системи	64
4.1 Дослідження впливу попередньої обробки даних та формування ознак	64
4.2 Порівняльний аналіз моделей машинного навчання.....	68
4.3 Дослідження впливу налаштування гіперпараметрів	72
4.4 Інтерпретація результатів прогнозування	74
4.5 Аналіз точності прогнозування	80
Висновки	86
Перелік джерел посилання	89
Додаток А Відомість кваліфікаційної роботи	93

ВСТУП

Оцінювання вартості об'єктів нерухомості є складною задачею, що відіграє важливу роль у багатьох сферах, зокрема при здійсненні операцій купівлі-продажу, інвестуванні, кредитуванні та управлінні активами. Формування ціни об'єкта залежить від значної кількості факторів, серед яких фізичні характеристики, розташування, технічний стан, рік побудови та інші параметри. Взаємодія цих факторів має складний та нелінійний характер, що ускладнює побудову універсальних моделей оцінювання вартості.

Традиційні підходи до визначення вартості нерухомості ґрунтуються на статистичних методах та експертних оцінках, що передбачають використання обмеженої кількості факторів і часто не враховують складні залежності між ними. У зв'язку з цим виникає необхідність застосування сучасних підходів, здатних ефективно працювати з великими обсягами даних та виявляти приховані закономірності.

Методи машинного навчання дозволяють будувати моделі, які автоматично визначають залежності між характеристиками об'єктів та їх вартістю. Використання таких методів відкриває можливість підвищення точності прогнозування, а також забезпечує гнучкість моделей при зміні структури даних. Окрім цього, сучасні підходи передбачають застосування методів інтерпретації результатів, що дає змогу аналізувати вплив окремих факторів на сформовану оцінку.

З огляду на зазначене, актуальним є дослідження та розробка інтелектуальних систем прогнозування вартості об'єктів на основі методів машинного навчання, які поєднують етапи обробки даних, побудови моделей та аналізу отриманих результатів. Такі системи можуть бути використані для автоматизованого оцінювання вартості об'єктів та підтримки процесів прийняття рішень.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Особливості формування вартості об'єктів нерухомості

Ринок нерухомості є складною та динамічною системою, у межах якої вартість об'єктів формується під впливом великої кількості взаємопов'язаних факторів. На відміну від більшості товарів, нерухомість має унікальні характеристики, тому навіть схожі об'єкти можуть суттєво відрізнятися за ринковою ціною. На формування вартості впливають фізичні параметри житла, місцезнаходження, економічна ситуація, рівень розвитку інфраструктури, ринкові тенденції та низка додаткових чинників [1].

Нерухомість одночасно виступає як товар ринку, інвестиційний актив та об'єкт довгострокового використання, що додатково ускладнює процес її оцінювання. На відміну від стандартизованих товарів, для яких існують чіткі критерії визначення вартості, ціна нерухомості значною мірою залежить від поєднання характеристик конкретного об'єкта та поточної ситуації на ринку. Навіть незначні відмінності у плануванні, стані будівлі, транспортній доступності або рівні розвитку району можуть суттєво впливати на ринкову вартість житла [2].

Особливістю ринку нерухомості є його висока чутливість до зовнішніх змін. Коливання економічних показників, зміни попиту, розвиток окремих районів або транспортної інфраструктури можуть суттєво впливати на ринкову вартість об'єктів навіть протягом короткого періоду часу. Через це задача оцінювання нерухомості є багатофакторною та потребує комплексного аналізу великого обсягу різномірних даних.

Додаткову складність створює динамічний характер ринку нерухомості. Вартість об'єктів може змінюватися залежно від економічної ситуації, сезонних тенденцій, інвестиційної активності та рівня попиту на

окремі типи житла. Крім того, значущість окремих факторів не є сталою та може змінюватися з часом. Наприклад, розвиток транспортної мережі або поява нової інфраструктури здатні суттєво підвищити привабливість певного району та вплинути на ринкову вартість нерухомості.

Основні фактори, що впливають на формування вартості об'єктів нерухомості, наведено на рисунку 1.1.



Рисунок 1.1 – Основні фактори формування вартості об'єктів нерухомості

Наведені фактори не існують ізольовано та часто взаємодіють між собою. Наприклад, недоліки планування можуть компенсуватися вигідним розташуванням, а сучасний житловий комплекс у віддаленому районі може мати вищу вартість порівняно зі старим житлом у центральній частині міста. Крім того, значущість окремих факторів може змінюватися залежно від поточної ситуації на ринку, типу нерухомості або потреб покупців. У

результаті оцінювання нерухомості не може базуватися лише на окремих характеристиках об'єкта та потребує комплексного підходу до аналізу [3].

Однією з ключових груп факторів є фізичні характеристики нерухомості. До них належать площа об'єкта, кількість кімнат, поверх, тип будівлі, рік побудови, технічний стан житла та особливості планування. Саме ці параметри найчастіше використовуються під час аналізу ринку нерухомості та формування вартості об'єктів.

У більшості випадків збільшення площі або покращення технічного стану житла позитивно впливає на його ринкову вартість. Водночас залежність між окремими характеристиками та ціною не є строго пропорційною. Наприклад, вартість одного квадратного метра може суттєво відрізнятися залежно від типу житла, планування або місцезнаходження об'єкта. Аналогічно однакові за площею квартири можуть мати різну ціну через стан будинку, поверх або якість ремонту.

Важливу роль також відіграють особливості внутрішнього планування та функціональність житлового простору. Рациональне використання площі, наявність окремих кімнат, просторої кухні або додаткових приміщень можуть підвищувати привабливість об'єкта для покупців. Водночас невдале планування або потреба у значних витратах на ремонт здатні знижувати ринкову вартість навіть за хорошого розташування об'єкта.

Одним із найважливіших факторів формування вартості нерухомості є місцезнаходження об'єкта. Локація значною мірою визначає рівень попиту на житло та безпосередньо впливає на його привабливість для потенційних покупців. Навіть за однакових фізичних характеристик нерухомість може суттєво відрізнятися за вартістю залежно від району розташування, транспортної доступності або рівня розвитку інфраструктури.

Основні характеристики місцезнаходження, що впливають на вартість нерухомості, наведено на рисунку 1.2.

Суттєве значення мають близькість до центру міста, транспортних вузлів, навчальних і медичних закладів, торгових центрів та зон відпочинку. Наявність розвиненої інфраструктури створює більш комфортні умови проживання та підвищує привабливість району для покупців. Особливо важливим цей фактор є для великих міст, де транспортна доступність та час пересування безпосередньо впливають на повсякденний комфорт мешканців.



Рисунок 1.2 – Основні характеристики місцезнаходження, що впливають на вартість нерухомості

На вартість нерухомості також впливають екологічна ситуація, рівень безпеки району, щільність забудови та загальна комфортність проживання. Наприклад, близькість до парків або зелених зон зазвичай позитивно впливає на ринкову ціну житла, тоді як високий рівень шуму або перевантаженість транспортної мережі можуть знижувати привабливість об'єкта.

Ринок нерухомості також залежить від економічної ситуації та загальних ринкових тенденцій. На формування вартості впливають рівень доходів населення, інфляція, валютний курс, доступність кредитування, інвестиційна активність та співвідношення попиту і пропозиції. У періоди економічного зростання попит на житло зазвичай збільшується, що призводить до підвищення цін, тоді як економічна нестабільність може спричиняти спад ринку.

Важливу роль відіграє також доступність іпотечного кредитування. Зниження кредитних ставок або спрощення умов отримання кредиту зазвичай стимулює попит на нерухомість, тоді як жорсткі умови кредитування можуть суттєво обмежувати активність ринку. Крім того, нерухомість часто розглядається як інвестиційний актив, тому зміни економічної ситуації безпосередньо впливають на інвестиційну активність та ринкову вартість об'єктів [4].

Додатковою особливістю ринку нерухомості є його динамічний характер. Вартість об'єктів змінюється з часом під впливом розвитку окремих районів, сезонних коливань попиту, змін інфраструктури та ринкової активності. Через це використання застарілих даних або статичних моделей оцінювання може призводити до суттєвих похибок прогнозування.

Традиційне визначення вартості нерухомості часто базується на експертному аналізі та порівнянні об'єкта з аналогічними пропозиціями на ринку. Проте такий підхід значною мірою залежить від досвіду оцінювача та повноти доступної інформації. Частина характеристик, зокрема комфортність району або якість планування, може оцінюватися суб'єктивно, що створює ризик похибок під час прогнозування вартості.

Додаткову складність створює наявність нелінійних залежностей між характеристиками нерухомості та її ринковою ціною. Вплив окремих параметрів може змінюватися залежно від типу об'єкта, локації або поточної ситуації на ринку. Через це класичні аналітичні підходи не завжди

дозволяють ефективно врахувати складну структуру взаємозв'язків між параметрами нерухомості.

У таких умовах перспективним напрямом є використання методів машинного навчання для прогнозування вартості нерухомості. На відміну від традиційних підходів, методи машинного навчання здатні аналізувати великі обсяги даних, автоматично виявляти приховані закономірності між характеристиками об'єктів та адаптуватися до змін ринку. Це дозволяє підвищити точність оцінювання та автоматизувати процес прогнозування вартості нерухомості [5].

1.2 Аналіз існуючих підходів і систем прогнозування вартості

Для початкових етапів розвитку автоматизованих систем оцінювання нерухомості характерним було використання класичних статистичних моделей, заснованих на аналізі залежності між характеристиками об'єкта та його ринковою вартістю. Найбільш поширеними підходами стали лінійна регресія, *log-linear regression* та *hedonic pricing model*, у межах яких вартість нерухомості розглядається як функція набору ознак, що описують фізичні характеристики житла, місцезнаходження та інші фактори.

Одним із прикладів застосування класичних регресійних підходів у задачі прогнозування вартості нерухомості є робота [6]. У дослідженні розглянуто задачу оцінювання вартості житлової нерухомості на основі гедонічної регресії та штучної нейронної мережі. Для аналізу використовувалися дані про характеристики житла, зокрема площу, кількість кімнат, тип будинку, тип будівлі, а також параметри, пов'язані з місцезнаходженням об'єкта.

У межах роботи гедонічна модель використовувалася для визначення залежності між характеристиками нерухомості та її ринковою вартістю. Результати аналізу показали, що найбільший вплив на формування ціни житла мають площа об'єкта, кількість кімнат, тип будинку, характеристики

місцезнаходження, тип будівлі, а також наявність додаткових зручностей, зокрема системи водопостачання та басейну. Отримані результати підтвердили, що вартість нерухомості формується під впливом сукупності фізичних і локаційних параметрів.

Для оцінювання ефективності підходу було виконано порівняння результатів гедонічної регресії з прогнозами, отриманими за допомогою штучної нейронної мережі. Результати прогнозування наведено на рисунку 1.3.

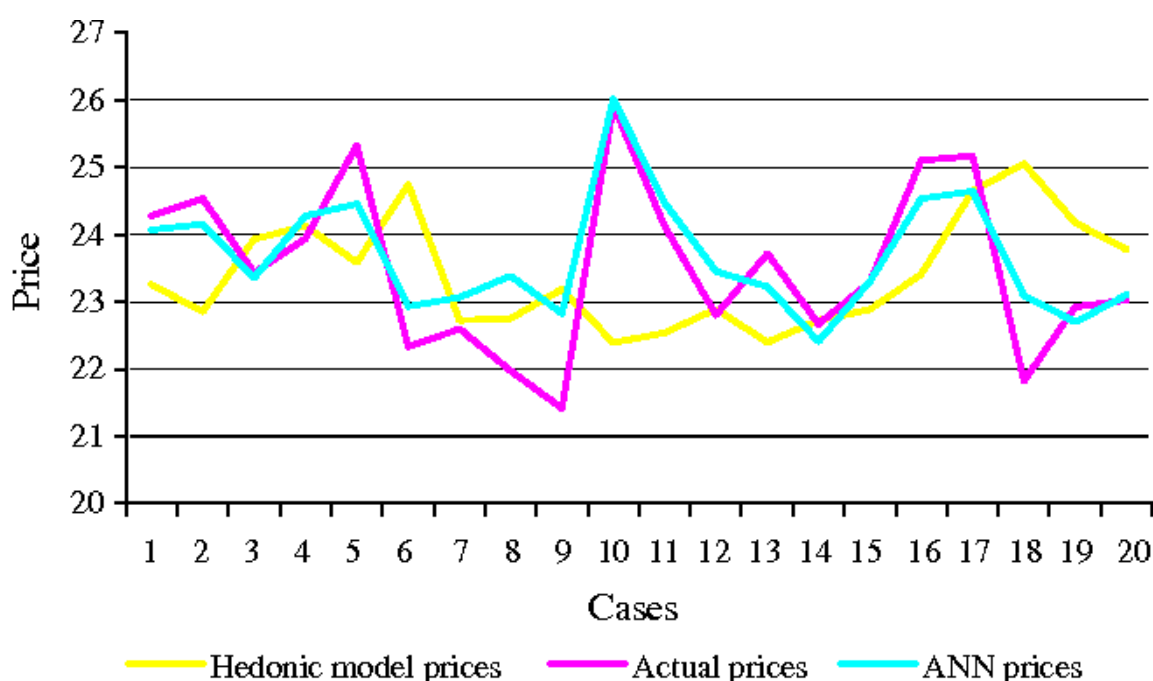


Рисунок 1.3 – Порівняння фактичних цін нерухомості з прогнозами гедонічної моделі та штучної нейронної мережі [6]

З рисунка видно, що прогнозовані значення, отримані за допомогою штучної нейронної мережі, у більшості випадків є ближчими до фактичних значень вартості нерухомості порівняно з результатами гедонічної регресії. Отримані результати демонструють обмеження класичних регресійних моделей у задачах прогнозування вартості нерухомості та підтверджують

доцільність використання методів машинного навчання для аналізу складних багатофакторних залежностей.

У роботі [7] досліджується застосування алгоритмів машинного навчання для прогнозування вартості житлової нерухомості на основі даних округу Fairfax County, Virginia. У межах дослідження автори виконали порівняння кількох моделей машинного навчання для аналізу залежності між характеристиками нерухомості та її ринковою вартістю.

Для побудови моделей використовувалися дані про характеристики житла, зокрема площу об'єкта, кількість кімнат, рік побудови, місцезнаходження та інші параметри, що впливають на формування ціни. У роботі застосовано алгоритми C4.5, RIPPER (Repeated Incremental Pruning to Produce Error Reduction), Naïve Bayesian та AdaBoost (Adaptive Boosting), які використовувалися для побудови моделей прогнозування вартості житла.

Для оцінювання ефективності моделей автори використовували метод 10-fold cross-validation та схему three-way split, у межах якої дані поділялися на навчальну, валідаційну та тестову вибірки. Результати порівняння моделей за рівнем похибки прогнозування наведено на рисунку 1.4.

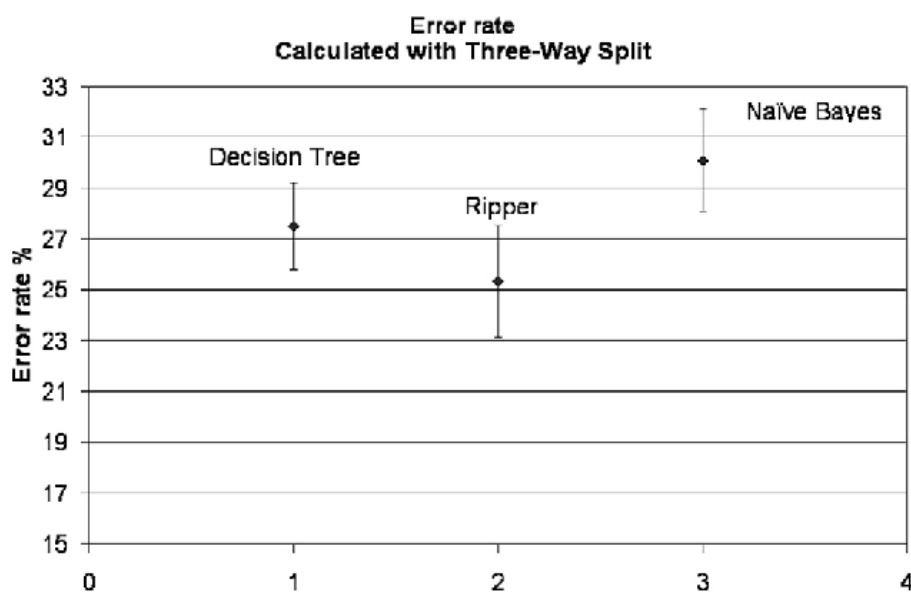


Рисунок 1.4 – Порівняння похибок прогнозування різних моделей машинного навчання [7]

З наведених результатів видно, що різні алгоритми демонструють різний рівень похибки прогнозування вартості житла. Найменшу похибку серед розглянутих підходів показали моделі RIPPER та AdaBoost, тоді як Naïve Bayesian продемонстрував нижчу точність прогнозування. Отримані результати підтверджують доцільність використання алгоритмів машинного навчання для задач оцінювання нерухомості та демонструють переваги ансамблевих підходів порівняно з більш простими моделями класифікації.

Подальший розвиток систем прогнозування вартості нерухомості пов'язаний із активним використанням алгоритмів машинного навчання та порівнянням їх ефективності у задачах оцінювання ринкової вартості житла. У сучасних дослідженнях найбільш поширеними підходами є Linear Regression, Random Forest, Gradient Boosting, XGBoost та CatBoost, які використовуються для аналізу складних залежностей між характеристиками нерухомості та її ціною. Значна частина робіт демонструє, що ансамблеві методи, засновані на поєднанні великої кількості дерев рішень, забезпечують вищу точність прогнозування порівняно з класичними регресійними моделями. Це пояснюється їх здатністю ефективно працювати з нелінійними залежностями, великою кількістю ознак та складною структурою ринкових даних.

У роботі [8] розглянуто задачу прогнозування вартості житла з використанням моделей Random Forest та XGBoost. Для експериментів використано набір даних Kaggle «House Prices – Advanced Regression Techniques», який містить 1420 зразків і 79 ознак, що описують характеристики житлових об'єктів. Перед побудовою моделей автор виконав попередню обробку даних, зокрема обробку пропущених значень, кореляційний аналіз і стандартизацію ознак.

У межах дослідження Random Forest та XGBoost були навчені й протестовані на одному наборі даних, що дозволило порівняти їхню ефективність у однакових умовах. За результатами експериментів XGBoost

показав вищу точність прогнозування та досяг значення коефіцієнта детермінації R^2 на рівні 89%. Автор зазначає, що обидві моделі можуть використовуватися для задачі прогнозування вартості житла, однак XGBoost забезпечив меншу похибку та кращу відповідність між фактичними й прогнозованими значеннями.

Приклад зіставлення фактичних і прогнозованих значень для моделі XGBoost наведено на рисунку 1.5.

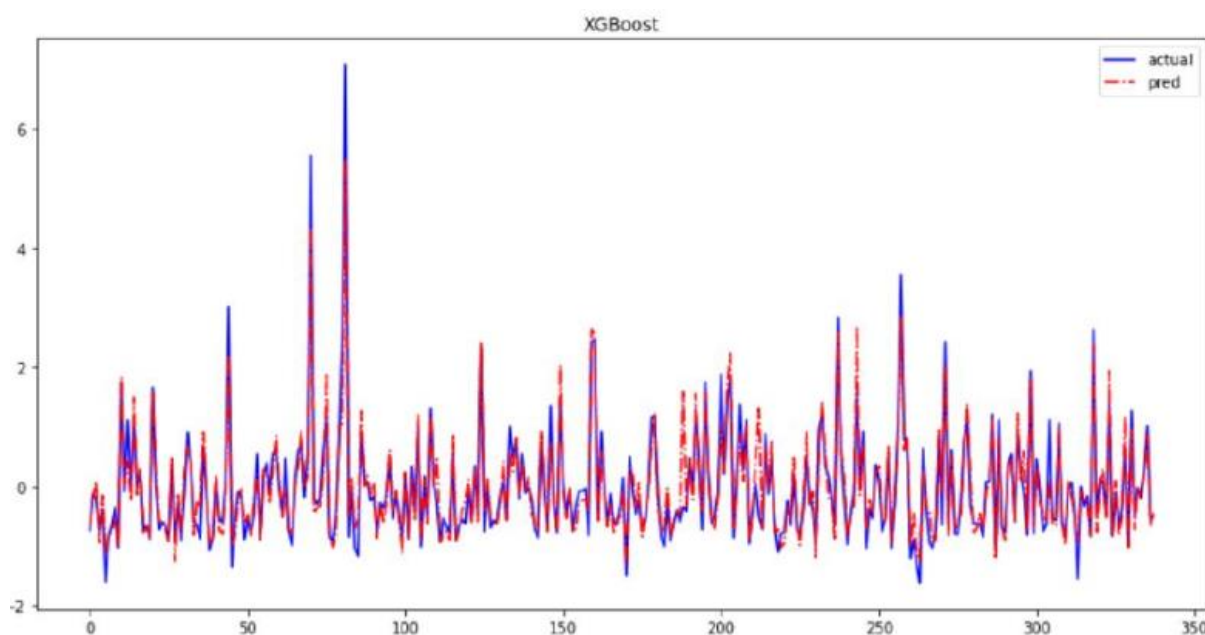


Рисунок 1.5 – Порівняння фактичних і прогнозованих значень вартості житла для моделі XGBoost [8]

Наведені результати демонструють ефективність градієнтного бустингу в задачі прогнозування вартості нерухомості. Водночас автор підкреслює, що отримані значення R^2 для обох моделей залишаються нижчими за 90%, тому такі моделі потребують подальшого вдосконалення перед практичним використанням. Як можливий напрям покращення запропоновано поєднання Random Forest і XGBoost, зокрема використання Random Forest для відбору найбільш значущих ознак перед навчанням моделі XGBoost.

У роботі [9] досліджується застосування алгоритмів машинного навчання для прогнозування вартості житлової нерухомості на прикладі ринку міста Цзінань, Китай. У межах дослідження автори виконали порівняння кількох моделей, зокрема Linear Regression, Random Forest та CatBoost, для визначення найбільш ефективного підходу до прогнозування цін на житло.

Для побудови моделей використовувалися дані про характеристики об'єктів нерухомості, що включали параметри житла, локаційні ознаки та додаткові фактори, пов'язані з ринковими умовами. Перед навчанням моделей виконувалася попередня обробка даних, зокрема очищення набору даних і підготовка ознак для подальшого аналізу.

Результати порівняння моделей за основними метриками якості наведено на рисунку 1.6.

	RMSE	R-square	MAE	Accuracy
Linear Regression	1973.054	0.779	30.681	78.45 %
Random Forest	948.810	0.893	18.129	89.77 %
Catboost	772.408	0.913	17.506	91.39 %

Рисунок 1.6 – Порівняння результатів моделей Linear Regression, Random Forest та CatBoost [9]

З наведених результатів видно, що ансамблеві методи Random Forest та CatBoost забезпечують вищу точність прогнозування порівняно з класичною лінійною регресією. Найкращі результати продемонструвала модель CatBoost, яка досягла значення R^2 на рівні 0.913 та точності прогнозування 91.39%. Крім того, модель CatBoost показала найменші значення RMSE та MAE, що свідчить про меншу похибку прогнозування вартості житла порівняно з іншими розглянутими підходами.

У роботі [10] досліджується задача прогнозування вартості житлової нерухомості з використанням різних алгоритмів машинного навчання. Для проведення експериментів автори використали набір даних Volusia County

Property Appraiser, що містить понад 62 тисячі записів про операції з нерухомістю за період 2015–2019 років. У роботі розглядалися моделі Linear Regression, Support Vector Regression, Decision Tree, Random Forest, XGBoost, CatBoost, Lasso та Voting Regressor.

Однією з особливостей дослідження є значна увага до етапу підготовки даних і формування ознак. Автори виконали очищення набору даних, видалення дублікатів, пропущених значень та викидів, а також аналіз кореляцій між ознаками. Для виявлення найбільш значущих параметрів використовувалися методи feature importance для моделей Random Forest та XGBoost. Крім того, у роботі запропоновано підхід target binning, який передбачає попередню категоризацію цільової змінної та подальше використання отриманих значень як додаткових ознак для навчання моделей.

Схему побудови системи прогнозування вартості нерухомості наведено на рисунку 1.7.

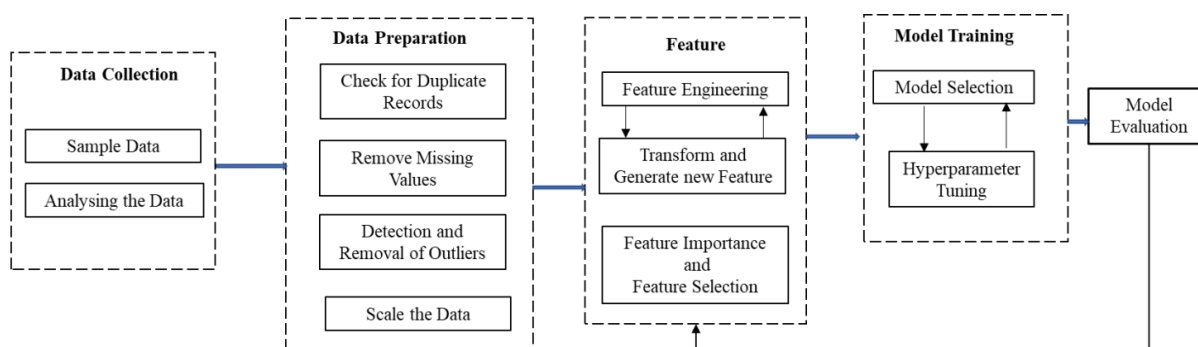


Рисунок 1.7 – Схема процесу побудови системи прогнозування вартості нерухомості [10]

Наведена схема демонструє основні етапи роботи системи, зокрема збір даних, попередню обробку, формування ознак, навчання моделей та оцінювання результатів прогнозування. Окрему увагу приділено feature

engineering і feature selection, що дозволило зменшити складність моделей та підвищити якість прогнозування.

У межах експериментів автори виконали порівняння різних алгоритмів машинного навчання за метриками R^2 , MSE та MAE. Результати показали, що ансамблеві методи забезпечують значно кращу точність прогнозування порівняно з класичними регресійними моделями. Зокрема, моделі XGBoost, CatBoost та Random Forest після застосування target binning досягли значення коефіцієнта детермінації R^2 на рівні 0.97.

На рисунку 1.8 наведено результати прогнозування для випадково вибраних тестових зразків. Видно, що моделі XGBoost, CatBoost та Voting Regressor формують прогнозовані значення, найбільш наближені до фактичної вартості нерухомості. Водночас класичні моделі, зокрема Linear Regression та Decision Tree, демонструють більші відхилення від реальних значень.

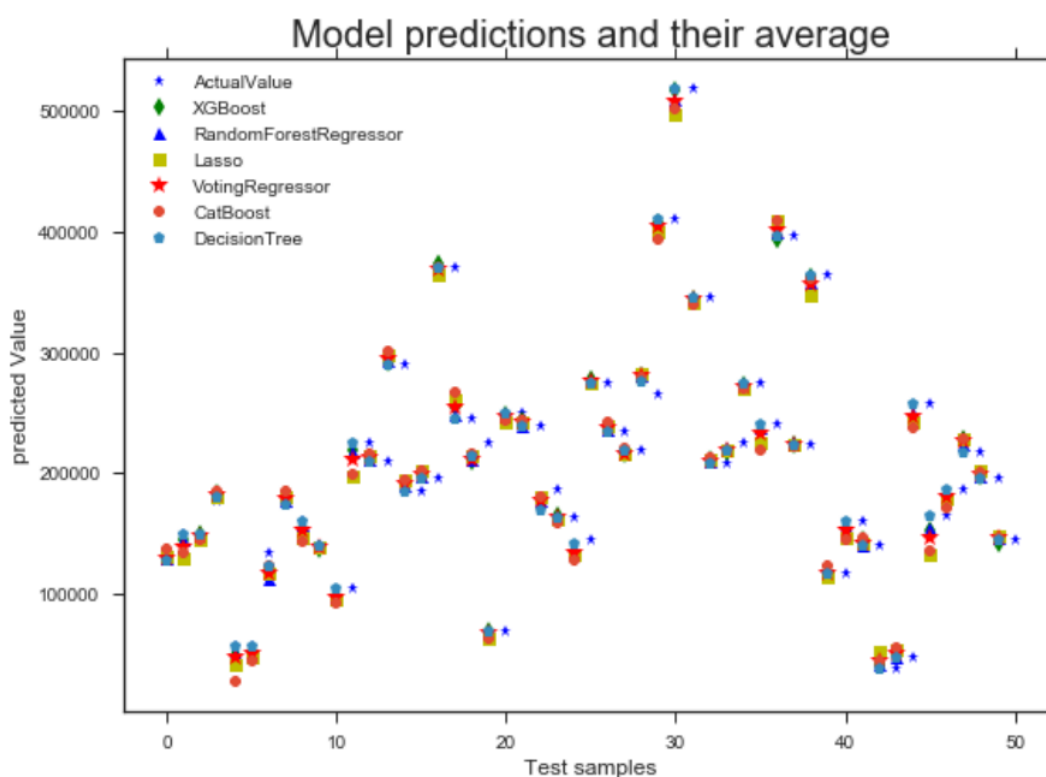


Рисунок 1.8 – Порівняння прогнозованих значень моделей із фактичними значеннями вартості нерухомості [10]

Автори також зазначають, що XGBoost забезпечив найкраще співвідношення між точністю прогнозування та швидкістю роботи моделі. Отримані результати підтверджують ефективність ансамблевих методів машинного навчання у задачах прогнозування вартості нерухомості та демонструють важливість попередньої обробки даних і формування інформативних ознак для підвищення якості моделей.

Окремим напрямом розвитку систем прогнозування вартості нерухомості є використання просторових та зовнішніх ознак, які дозволяють враховувати особливості розташування об'єкта та характеристики навколишнього середовища. У таких підходах для побудови моделей використовуються не лише параметри житла, а й інформація про район розташування, транспортну доступність, відстані до об'єктів інфраструктури, рівень безпеки, екологічну ситуацію та соціально-економічні показники території. Використання просторових ознак дозволяє точніше враховувати вплив місцезнаходження на формування ринкової вартості нерухомості та підвищувати якість прогнозування у складних міських умовах.

У роботі [11] досліджується вплив просторової залежності на точність прогнозування вартості житлової нерухомості. Автори зазначають, що класичні гедонічні моделі не завжди здатні коректно враховувати просторові взаємозв'язки між об'єктами нерухомості, оскільки сусідні об'єкти часто мають схожі характеристики та спільні локаційні фактори. У межах дослідження виконано порівняння кількох підходів до врахування *spatial dependence*, зокрема OLS-моделей, *geostatistical models*, *trend surface models* та моделей із використанням *nearest neighbors residuals*.

Для експериментів використовувалися дані приблизно про 13 тисяч угод купівлі-продажу житла в Louisville, Kentucky. Під час побудови моделей враховувалися не лише фізичні характеристики житла, а й просторові параметри, зокрема координати розташування, належність до окремих районів і характеристики локальних субринків нерухомості.

Автори також виконали кластеризацію територій та формування transaction groups для більш точного врахування просторових особливостей ринку.

Результати експериментів показали, що врахування просторової залежності дозволяє суттєво підвищити точність прогнозування вартості нерухомості. Найкращі результати продемонструвала geostatistical model із використанням детально сформованих spatial submarkets. Зокрема, точність прогнозування для geostatistical model із 60 transaction groups досягла 47.4% прогнозів у межах 10% від фактичної ціни, тоді як базова OLS-модель без урахування просторових факторів забезпечила лише 36.5%. Отримані результати підтверджують важливість використання просторових та локаційних ознак у системах прогнозування вартості нерухомості.

У роботі [12] запропоновано підхід до прогнозування вартості нерухомості з урахуванням просторово-часових залежностей між об'єктами. Дослідження базується на аналізі понад 428 тисяч угод купівлі-продажу житла в місті Adelaide (Австралія) за період 1984–2016 років. Для побудови моделей використовувалися 38 пояснювальних змінних, які охоплювали фізичні характеристики житла, показники транспортної доступності, параметри навколишнього середовища та соціально-економічні характеристики районів.

У роботі було реалізовано кілька моделей машинного навчання, серед яких Linear Regression, Decision Tree, Random Forest та Gradient-Boosted Trees. Особливістю дослідження стало використання spatio-temporal lag variable – спеціальної ознаки, що враховує просторову та часову близькість сусідніх об'єктів нерухомості. Автори зазначають, що врахування spatio-temporal dependency дозволяє підвищити точність прогнозування вартості житла порівняно з моделями, які використовують лише стандартні характеристики об'єктів.

На рисунку 1.9 наведено загальну схему процесу прогнозування вартості нерухомості, запропонованого авторами. Схема демонструє етапи

збору та інтеграції даних, формування ознак, попередньої обробки даних, побудови моделей машинного навчання та оцінювання їх точності.

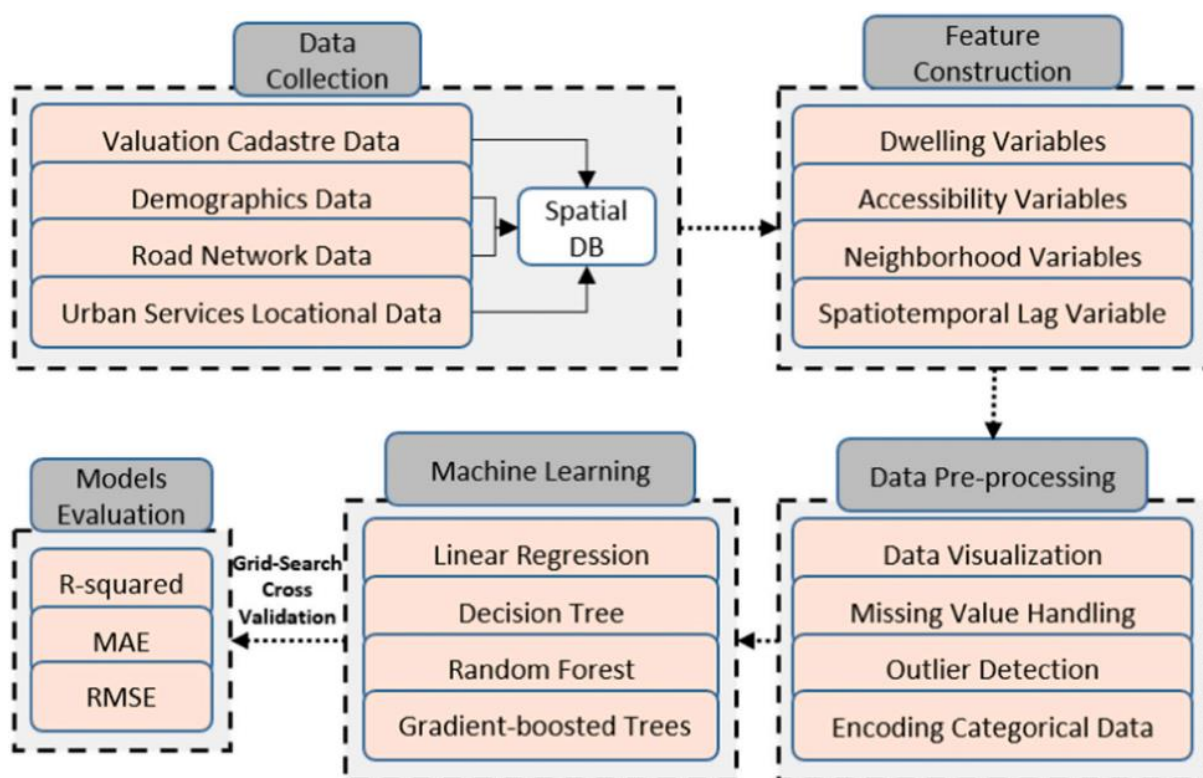


Рисунок 1.9 – Схема процесу прогнозування вартості нерухомості з урахуванням просторово-часових залежностей [12]

Під час підготовки даних автори виконували обробку пропущених значень, виявлення викидів, нормалізацію даних та кодування категоріальних ознак. Також досліджувалась кореляція між параметрами нерухомості та ринковою вартістю об'єктів. Особливу увагу було приділено формуванню spatial-temporal lag feature, яка враховувала характеристики сусідніх продажів у межах визначеного просторового та часового радіуса.

Результати експериментів показали, що ансамблеві моделі забезпечують вищу точність прогнозування порівняно з лінійною регресією та окремими деревами рішень. Найкращі результати продемонструвала модель Gradient-Boosted Trees, для якої було отримано значення $R^2 = 0.896$, $RMSE = 0.086$ та $MAE = 0.058$. Автори також встановили, що використання

spatio-temporal lag variable суттєво покращує якість прогнозування для всіх досліджуваних моделей.

На рисунку 1.10 наведено результати оцінювання важливості ознак у моделі Gradient-Boosted Trees. Як показано на рисунку, найбільший вплив на прогнозовану вартість мали spatio-temporal lag, площа житла, рівень освіти та зайнятості населення району, вік будівлі та відстань до транспортної інфраструктури.

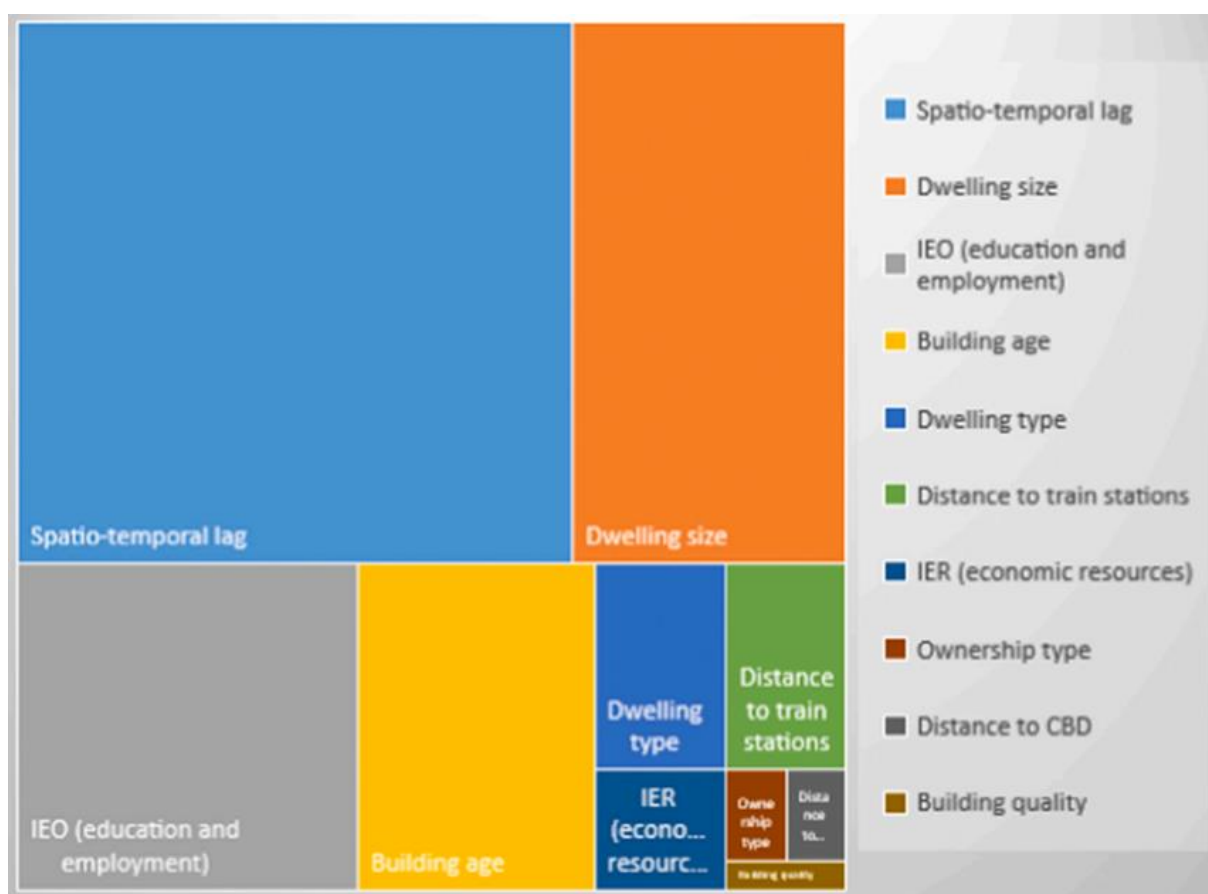


Рисунок 1.10 – Найважливіші ознаки моделі Gradient-Boosted Trees [12]

Отримані результати підтверджують, що врахування просторово-часових залежностей та використання ансамблевих методів машинного навчання дозволяє суттєво підвищити точність прогнозування вартості нерухомості. Робота також демонструє важливість інтеграції географічних,

соціально-економічних та часових параметрів у сучасних системах оцінювання нерухомості.

Сучасні системи прогнозування вартості нерухомості дедалі частіше використовують ансамблеві підходи, що поєднують декілька моделей машинного навчання. Одним із таких підходів є *stacked ensemble*, у якому результати кількох базових моделей передаються до метамоделі, що формує фінальний прогноз. Використання подібних підходів дозволяє врахувати переваги різних алгоритмів та підвищити точність оцінювання вартості нерухомості.

У роботі [13] досліджено застосування *stacked ensemble* підходу для автоматизованого оцінювання житлової нерухомості. Для побудови моделі використовувався набір даних, що містив понад 63 тисячі записів про операції з нерухомістю на острові Jeju за період 2012–2021 років. У межах дослідження враховувалися характеристики об'єктів нерухомості, параметри житлових комплексів та просторові ознаки, зокрема координати, інформація про школи та зонування території.

Автори реалізували кілька моделей машинного навчання, серед яких *k*-NN, Decision Tree, Random Forest, Gradient Boosting, XGBoost, CatBoost та AdaBoost. Для формування фінального прогнозу використовувався *stacked ensemble* підхід, у якому результати базових моделей передавалися до метамоделі на основі Ridge Regression. Для оцінювання якості прогнозування використовувалися метрики RMSE, MAE, MAPE, CVRMSE та коефіцієнт детермінації R^2 . За результатами експериментів *stacked ensemble* модель продемонструвала найкращі показники серед усіх досліджуваних методів, досягнувши $R^2 = 0.9725$ та найменших значень похибок прогнозування.

Окрім оцінювання точності прогнозування, у роботі було проведено додатковий аналіз умов, за яких модель демонструє зниження якості прогнозування. Для цього автори використали *decision tree analysis*, що дозволив визначити характеристики об'єктів нерухомості, пов'язані з

підвищенням похибки прогнозування. Аналіз показав, що найбільші труднощі для моделі виникають під час оцінювання невеликих та старих житлових комплексів. Було встановлено, що саме розмір житлового комплексу, наявність ліфта та вік будівлі є ключовими факторами, які впливають на зниження точності прогнозування. Результати decision tree analysis наведено на рисунку 1.11.

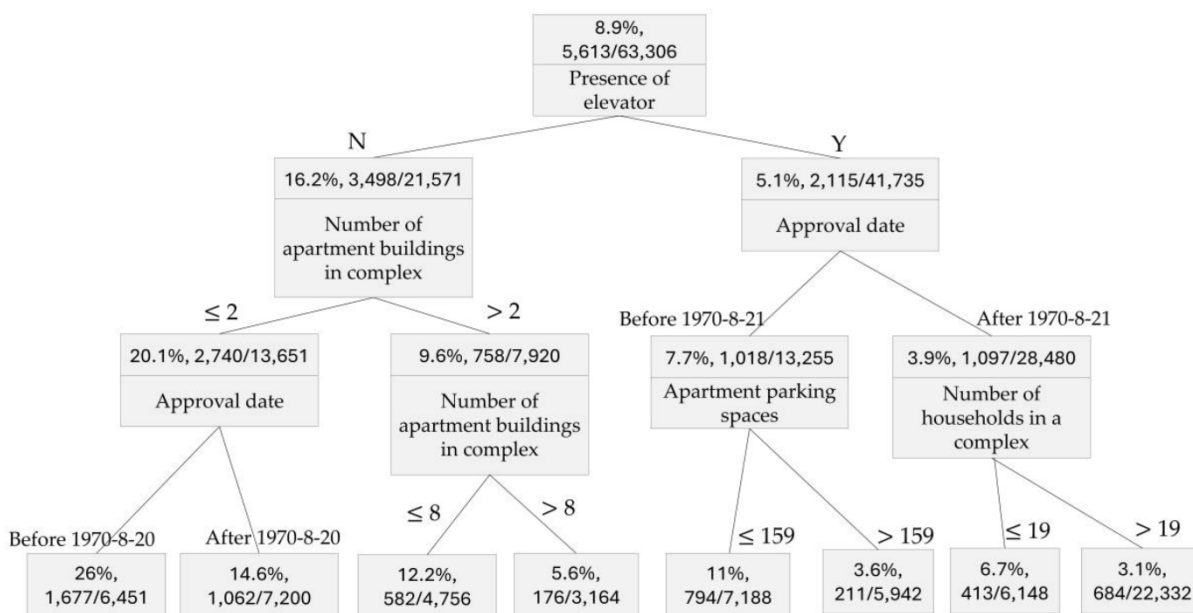


Рисунок 1.11 – Decision tree analysis факторів, що впливають на зниження точності прогнозування stacked ensemble моделі [13]

Як видно з рисунка, найбільші значення похибки прогнозування характерні для невеликих житлових комплексів без ліфта та старих будівель, введених в експлуатацію до 1970 року. Автори зазначають, що для таких об'єктів можуть бути необхідні додаткові методи аналізу або використання більш спеціалізованих моделей оцінювання. Отримані результати демонструють, що складні ансамблеві моделі дозволяють суттєво підвищити точність прогнозування вартості нерухомості, однак їх практичне використання потребує додаткового аналізу умов, за яких точність моделей може знижуватися.

Останніми роками значного поширення набули explainable AI підходи, спрямовані не лише на підвищення точності моделей машинного навчання, а й на забезпечення прозорості їх роботи. Для задач прогнозування вартості нерухомості це має особливе значення, оскільки результати оцінювання часто використовуються під час прийняття фінансових та інвестиційних рішень. У зв'язку з цим сучасні дослідження дедалі частіше поєднують ансамблеві методи машинного навчання з інструментами інтерпретації моделей, такими як SHAP, LIME, PDP та ALE. Це дозволяє не лише формувати точні прогнози, а й визначати ознаки, які найбільше впливають на результати оцінювання нерухомості.

У роботі [14] запропоновано підхід до побудови систем прогнозування вартості нерухомості, який поєднує ансамблеві моделі машинного навчання з методами explainable AI та fairness-aware learning. Основна увага у дослідженні приділяється не лише точності прогнозування, а й забезпеченню прозорості роботи моделей, можливості пояснення отриманих результатів та аналізу потенційних упереджень під час прийняття рішень.

Для прогнозування вартості нерухомості автори використовували моделі LightGBM та XGBoost. Під час побудови моделей враховувалися характеристики житла, площа будинку, якість будівництва, рік побудови, площа житлових приміщень, площа підвалу, параметри земельної ділянки та інші характеристики об'єктів нерухомості. Перед навчанням моделей виконувалася попередня обробка даних, що включала нормалізацію числових ознак, feature engineering, кодування категоріальних параметрів та оптимізацію гіперпараметрів моделей за допомогою grid search.

Особливістю роботи є комплексне використання explainable AI методів для аналізу результатів прогнозування. Для пояснення роботи моделей застосовувалися SHAP summary plots, SHAP force plots, Partial Dependence Plots (PDP), Accumulated Local Effects (ALE) та LIME explanations. Використання кількох підходів одночасно дозволило авторам

виконати як глобальний аналіз важливості ознак, так і локальне пояснення окремих прогнозів моделей. Автори зазначають, що подібні інструменти є особливо важливими для фінансових та ринкових систем, де необхідно забезпечити прозорість процесу прогнозування та можливість інтерпретації результатів для кінцевих користувачів.

На рисунку 1.12 наведено SHAP summary plot для моделі LightGBM. Як видно з рисунка, найбільший вплив на результати прогнозування мають ознаки OverallQual, GrLivArea, TotalBsmtSF та YearBuilt. Крім визначення важливості ознак, SHAP summary plot дозволяє оцінити напрямок їх впливу на прогнозовану вартість об'єкта. Зокрема, високі значення OverallQual, GrLivArea та TotalBsmtSF переважно призводять до збільшення прогнозованої вартості нерухомості, тоді як нижчі значення цих параметрів зменшують прогноз моделі. Таким чином, SHAP summary plot дозволяє одночасно оцінити як значущість ознак, так і характер їх впливу на результат прогнозування.

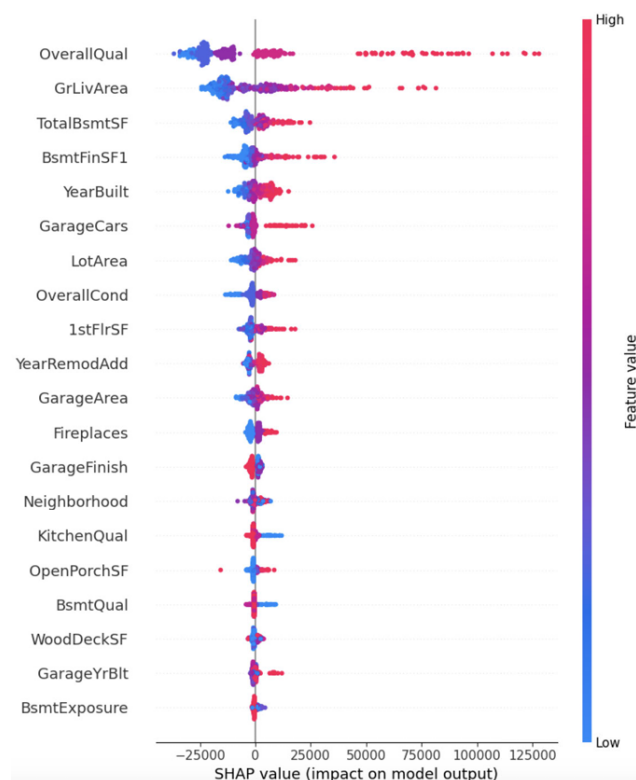


Рисунок 1.12 – SHAP summary plot для моделі LightGBM [14]

Для локального пояснення окремих прогнозів автори використовували SHAP force plots. Приклад пояснення прогнозу для конкретного об'єкта нерухомості наведено на рисунку 1.13.



Рисунок 1.13 – SHAP force plot для окремого прогнозу моделі LightGBM [14]

SHAP force plot демонструє внесок окремих ознак у формування фінального прогнозу моделі. На рисунку показано, які характеристики збільшують прогнозовану вартість об'єкта, а які, навпаки, знижують її. Подібний підхід дозволяє пояснювати причини отримання конкретного результату прогнозування та робить процес роботи ансамблевої моделі більш прозорим для користувача. Автори зазначають, що локальні SHAP-пояснення можуть використовуватися для аналізу помилок прогнозування та виявлення випадків, у яких модель формує нестабільні результати.

Додатково у роботі використовувався метод LIME для побудови локальних пояснень прогнозів моделі. Приклад LIME explanation наведено на рисунку 1.14.

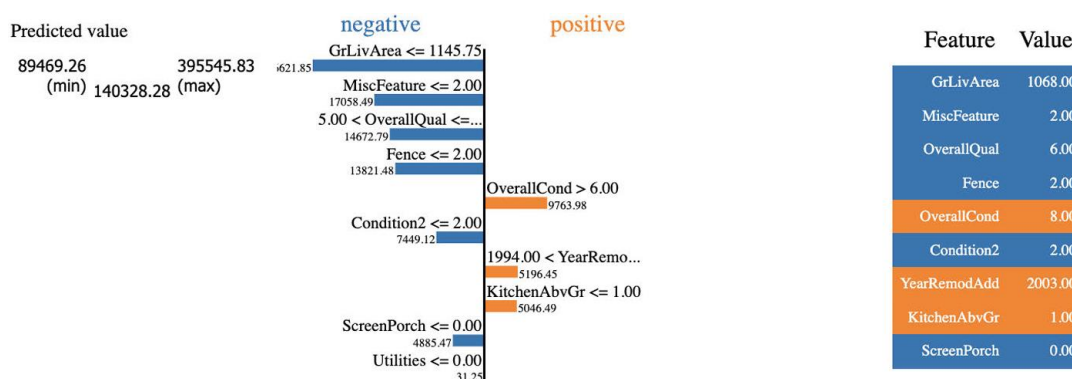


Рисунок 1.14 – LIME explanation для моделі LightGBM [14]

На відміну від SHAP, метод LIME формує локальну інтерпретовану модель для окремого прогнозу та дозволяє оцінити внесок ознак у більш спрощеному вигляді. LIME explanation демонструє ознаки, які найбільше вплинули на формування конкретного прогнозу, а також напрямок їх впливу. Автори зазначають, що використання LIME є корисним для пояснення результатів роботи складних ансамблевих моделей користувачам, які не мають спеціалізованих знань у сфері машинного навчання.

У роботі також проводився аналіз fairness-aware learning, у межах якого досліджувався вплив fairness constraints на точність та інтерпретованість моделей. Автори встановили, що додавання fairness constraints може дещо знижувати точність прогнозування, однак дозволяє зменшити ризик формування упереджених рішень щодо окремих груп об'єктів або користувачів. За результатами експериментів модель LightGBM забезпечила краще співвідношення між точністю прогнозування, fairness metrics та explainability порівняно з XGBoost. Отримані результати підтверджують, що сучасні системи прогнозування вартості нерухомості повинні враховувати не лише точність моделей, а й прозорість їх роботи та можливість інтерпретації результатів прогнозування.

1.3 Постановка задачі дослідження

Проведений аналіз предметної галузі показав, що формування вартості об'єктів нерухомості залежить від значної кількості взаємопов'язаних факторів, серед яких площа, розташування, технічний стан, характеристики будівлі та інші параметри. При цьому залежності між ознаками та вартістю об'єкта мають складний і нелінійний характер, що ускладнює використання традиційних підходів до оцінювання.

Розгляд існуючих підходів і систем прогнозування вартості свідчить про ефективність сучасних методів машинного навчання у задачах

прогнозування завдяки здатності працювати з великою кількістю ознак та виявляти приховані закономірності у даних. Особливу результативність демонструють ансамблеві моделі та методи градієнтного бустингу. Водночас важливим аспектом є не лише отримання точного прогнозу, а й можливість інтерпретації результатів моделі та визначення впливу окремих факторів на сформовану оцінку.

У зв'язку з цим актуальною є задача розробки інтелектуальної системи прогнозування вартості об'єктів на основі методів машинного навчання, яка забезпечує повний цикл обробки даних, побудови моделей, оцінювання їх якості та аналізу отриманих результатів.

Метою роботи є розробка інтелектуальної системи прогнозування вартості об'єктів на основі методів машинного навчання.

Для досягнення поставленої мети необхідно виконати такі задачі:

- проаналізувати особливості предметної галузі та фактори, що впливають на формування вартості об'єктів;
- дослідити існуючі підходи та системи прогнозування вартості;
- проаналізувати методи машинного навчання, що застосовуються для задач прогнозування;
- підготувати та обробити набір даних для побудови моделей;
- реалізувати моделі прогнозування вартості об'єктів на основі методів машинного навчання;
- провести оцінювання та порівняльний аналіз якості моделей;
- здійснити аналіз впливу окремих ознак на прогнозовану вартість об'єктів.

2 МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВАРТОСТІ ОБ'ЄКТІВ

Ефективність систем прогнозування вартості об'єктів значною мірою залежить від вибору методів обробки та аналізу даних. Традиційні статистичні підходи, що тривалий час використовувалися для оцінювання вартості, мають обмежені можливості щодо виявлення складних нелінійних залежностей між характеристиками об'єкта та його ринковою ціною. У зв'язку з цим все більшого поширення набувають методи машинного навчання, здатні автоматично виявляти приховані закономірності у великих масивах даних та формувати більш точні прогнози.

Сучасні алгоритми машинного навчання охоплюють широкий спектр підходів, серед яких особливе місце займають регресійні та ансамблеві методи. Такі моделі дозволяють враховувати одночасний вплив великої кількості ознак, адаптуватися до складної структури даних та забезпечувати високу якість прогнозування. Крім того, важливим аспектом є можливість оцінювання точності побудованих моделей та інтерпретації отриманих результатів, що підвищує довіру до прогнозів і сприяє їх практичному використанню.

2.1 Регресійні методи машинного навчання

Регресійні методи машинного навчання призначені для розв'язання задач, у яких необхідно передбачити числове значення цільової змінної на основі набору вхідних ознак. На відміну від класифікації, результатом якої є належність об'єкта до одного з наперед визначених класів, регресія формує безперервний прогноз, наприклад вартості нерухомості, автомобіля чи іншого об'єкта. При цьому регресійні та класифікаційні алгоритми не є взаємовиключними підходами, а часто використовуються спільно в межах складних інтелектуальних систем, де класифікація може визначати

подальший сценарій обробки даних, а регресія – виконувати прогнозування кількісних показників [15].

Одним із найпростіших і водночас фундаментальних методів регресійного аналізу є лінійна регресія. Її основна ідея полягає у знаходженні такої прямої, яка найкраще описує залежність між незалежними змінними та цільовим показником, мінімізуючи різницю між фактичними та прогнозованими значеннями [16]. Отримана модель може бути використана для прогнозування значення цільової змінної для нових об'єктів, характеристики яких відомі.

Принцип роботи лінійної регресії наведено на рисунку 2.1. На ньому показано набір експериментальних даних, лінію регресії, прогнозовані значення та залишки (помилки), які характеризують відхилення фактичних значень від побудованої моделі.

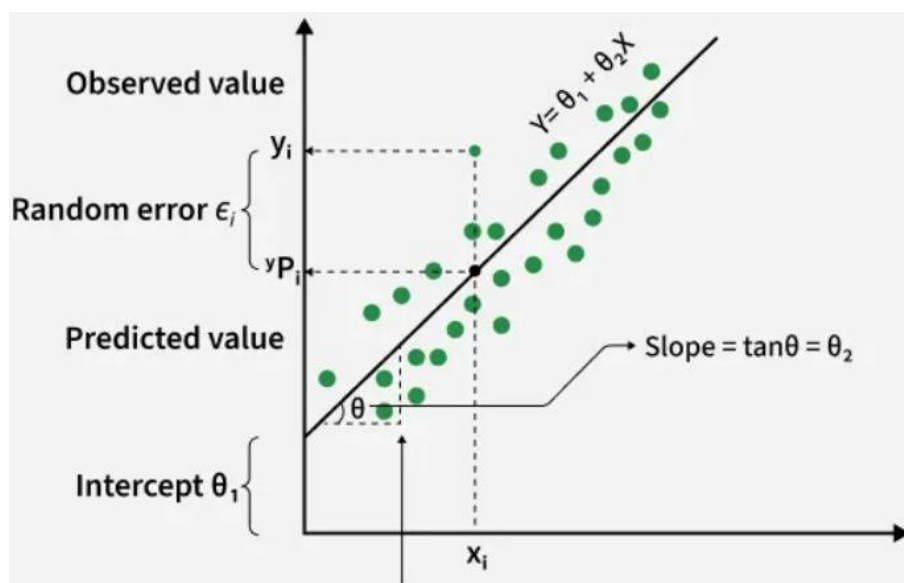


Рисунок 2.1 – Принцип роботи моделі лінійної регресії [16]

Як видно з рисунка 2.1, лінія регресії описується рівнянням, у якому залежна змінна Y визначається значенням незалежної змінної X . При цьому X виступає предиктором, тобто ознакою, на основі якої здійснюється прогнозування, а Y – цільовою змінною, значення якої необхідно

передбачити. Коефіцієнт θ_1 визначає точку перетину прямої з віссю ординат і відповідає значенню прогнозу при $X = 0$, тоді як коефіцієнт θ_2 характеризує нахил прямої та показує, як змінюється прогнозоване значення Y при збільшенні X на одну одиницю. Різниця між фактичним і прогнозованим значеннями утворює залишок (помилку), мінімізація якого є основною метою навчання моделі.

Незважаючи на простоту, лінійна регресія є базовою моделлю регресійного аналізу та широко використовується як відправна точка під час дослідження ефективності більш складних алгоритмів машинного навчання. Водночас на практиці задачі прогнозування часто залежать не від одного, а від сукупності факторів. У такому випадку застосовується множинна лінійна регресія, яка враховує одночасний вплив кількох незалежних змінних на значення цільової ознаки. Модель множинної лінійної регресії описується таким рівнянням:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n, \quad (2.1)$$

де $\hat{y}$ – прогнозоване значення цільової змінної;

$x_1, x_2, \dots, x_n$ – незалежні змінні (ознаки);

$\theta_1, \theta_2, \dots, \theta_n$ – вагові коефіцієнти, що визначають внесок кожної ознаки у формування прогнозу;

θ_0 – вільний член моделі.

Основною метою множинної лінійної регресії є знаходження такої комбінації коефіцієнтів, яка забезпечує найточніше прогнозування цільової змінної на основі наявного набору ознак. Саме тому цей підхід широко застосовується в задачах прогнозування вартості нерухомості, фінансового аналізу, оцінювання врожайності, прогнозування продажів та інших прикладних задачах, де результат залежить від одночасного впливу багатьох факторів [16].

На відміну від лінійної регресії, яка описує залежність між ознаками та цільовою змінною за допомогою єдиного рівняння, дерева рішень формують прогноз шляхом послідовного розбиття множини даних на більш однорідні підмножини. Такий підхід дозволяє моделі враховувати складні нелінійні залежності між ознаками без необхідності попереднього задання математичної форми зв'язку.

Структура дерева рішень складається з кореневого вузла, внутрішніх вузлів та листків. Кореневий вузол містить усі навчальні дані, після чого виконується їх послідовний поділ за певними умовами. Кожний внутрішній вузол відповідає перевірці значення однієї з ознак, а гілки визначають можливі результати цієї перевірки. Кінцеві вузли (листки) містять прогнозоване значення цільової змінної, яке використовується під час оцінювання нового об'єкта. Загальний принцип побудови дерева рішень наведено на рисунку 2.2.

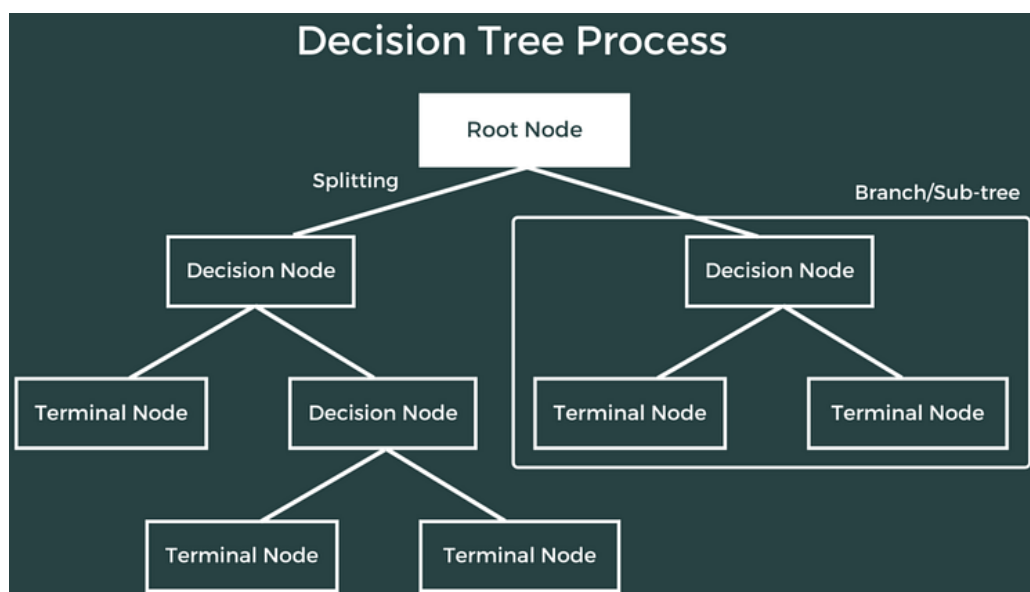


Рисунок 2.2 – Загальна структура дерева рішень [17]

Побудова дерева здійснюється шляхом рекурсивного двійкового розбиття навчальної вибірки. На кожному вузлі алгоритм аналізує всі можливі варіанти поділу даних та обирає той, який забезпечує найкраще

відокремлення спостережень. У задачах регресії критерієм якості такого розбиття найчастіше виступає середньоквадратична помилка (Mean Squared Error, MSE). Для кожного можливого поділу обчислюється значення MSE окремо для лівої та правої підмножин, після чого визначається їх зважене середнє. Найкращим вважається той поділ, який забезпечує мінімальне значення цієї величини.

Після виконання розбиття аналогічна процедура повторюється для кожної отриманої підмножини. Процес триває рекурсивно доти, доки не буде досягнуто одного або кількох критеріїв зупинки. До таких критеріїв належать максимальна глибина дерева, мінімальна кількість об'єктів у вузлі або недостатнє зменшення помилки після чергового розбиття [17]. Використання критеріїв зупинки дозволяє обмежити складність моделі та зменшити ризик її перенавчання.

2.2 Ансамблеві методи машинного навчання

Незважаючи на простоту та зрозумілість дерев рішень, використання одного дерева не завжди забезпечує достатню точність прогнозування. Такі моделі є чутливими до змін навчальної вибірки: навіть незначні відмінності у вхідних даних можуть призвести до побудови іншої структури дерева та зміни результатів прогнозування. Крім того, окремі дерева часто схильні до перенавчання, що негативно впливає на їх здатність узагальнювати знання на нових даних.

Для усунення цих недоліків були розроблені ансамблеві методи машинного навчання. Їх основна ідея полягає не у використанні однієї моделі, а у поєднанні результатів роботи кількох базових моделей. Завдяки цьому вдається підвищити стійкість алгоритму до випадкових коливань даних, зменшити похибку прогнозування та покращити здатність моделі працювати з новими об'єктами. У більшості випадків базовими моделями

ансамблів виступають дерева рішень, які навчаються за певною схемою та формують спільний прогноз.

Залежно від способу побудови ансамблю методи ансамблевого навчання поділяють на дві основні групи: паралельні та послідовні. У паралельних ансамблях кожна базова модель навчається незалежно від інших, після чого їх результати об'єднуються для формування остаточного прогнозу. Такий підхід дозволяє зменшити дисперсію моделі та підвищити її стійкість до перенавчання. Натомість у послідовних ансамблях кожна наступна модель навчається з урахуванням помилок, допущених попередніми моделями. У результаті кожний новий етап навчання спрямований на поступове підвищення точності прогнозування шляхом коригування вже побудованої моделі [18].

Найбільш поширеними представниками цих підходів є bagging (Bootstrap Aggregating), який належить до паралельних ансамблевих методів, та boosting, що реалізує принцип послідовного навчання моделей. Схематичне порівняння цих двох підходів наведено на рисунку 2.3.

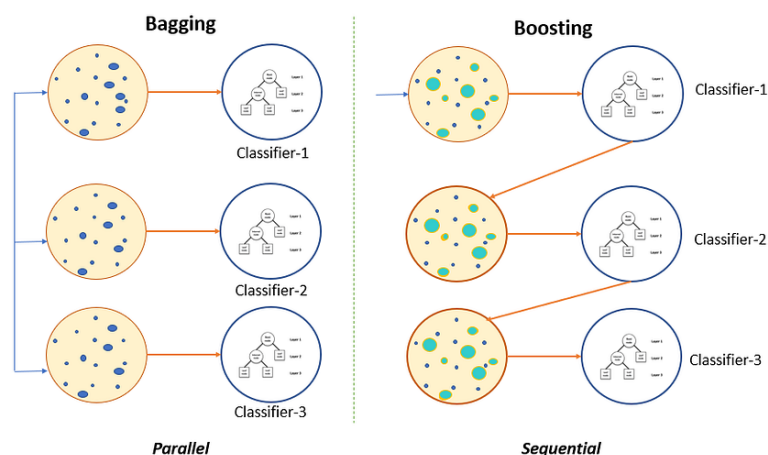


Рисунок 2.3 – Порівняння підходів bagging та boosting [19]

Метод bagging передбачає незалежне навчання кількох базових моделей на різних підвибірках, сформованих із початкового набору даних за допомогою бутстреп-методу. Усі моделі будуються паралельно та не

взаємодіють між собою під час навчання, тому помилки окремих моделей не впливають на процес побудови інших. Після завершення навчання результати роботи всіх моделей об'єднуються, що дозволяє отримати більш стабільний та точний прогноз порівняно з використанням одного дерева рішень.

На відміну від bagging, метод boosting реалізує послідовний підхід до побудови ансамблю. Кожна наступна модель навчається після попередньої та враховує помилки, яких вона припустилася. Таким чином, основною метою кожного нового етапу навчання є зменшення залишкової похибки ансамблю. У багатьох алгоритмах бустингу особлива увага приділяється саме тим спостереженням, для яких попередні моделі сформували найбільшу помилку прогнозування [19].

На відміну від bagging, де всі навчальні приклади мають однаковий вплив на побудову моделей, у boosting окремим спостереженням можуть призначатися різні ваги. Об'єкти, що були класифіковані або спрогнозовані з більшою похибкою, отримують вищу вагу на наступних етапах навчання, тоді як для добре описаних спостережень їхній вплив поступово зменшується. Такий механізм дозволяє ансамблю концентруватися на найбільш складних прикладах та забезпечує вищу точність прогнозування.

Однією з найпоширеніших реалізацій підходу bagging є алгоритм Random Forest (випадковий ліс), який являє собою ансамбль дерев рішень. На відміну від окремого дерева, Random Forest одночасно будує велику кількість незалежних моделей, після чого об'єднує їх результати для формування остаточного прогнозу. Завдяки цьому вдається істотно зменшити вплив випадкових коливань навчальної вибірки, підвищити стійкість моделі та покращити точність прогнозування.

Кожне дерево ансамблю навчається не на всій навчальній вибірці, а на окремій бутстреп-вибірці, сформованій випадковим чином із повтореннями. При цьому частина початкових спостережень не потрапляє до конкретної вибірки та утворює так званий out-of-bag набір, який може

використовуватися для внутрішнього оцінювання якості моделі без необхідності формування окремої тестової вибірки [20].

Додатковий рівень випадковості досягається завдяки випадковому відбору ознак під час побудови кожного вузла дерева. Замість аналізу всіх доступних характеристик алгоритм розглядає лише випадково вибрану їх підмножину. Це зменшує кореляцію між окремими деревами, підвищує різноманітність ансамблю та сприяє покращенню його узагальнювальної здатності.

Принцип роботи алгоритму Random Forest наведено на рисунку 2.4. Як видно з рисунка, кілька незалежно побудованих дерев рішень формують власні прогнози, які надалі агрегуються для отримання кінцевого результату. У задачах регресії остаточне прогнозоване значення визначається як середнє арифметичне прогнозів усіх дерев ансамблю, тоді як у задачах класифікації використовується принцип більшості голосів.

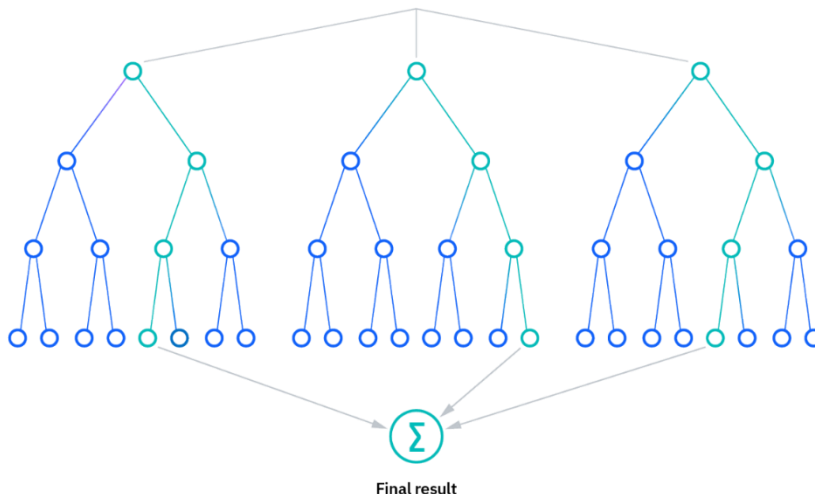


Рисунок 2.4 – Принцип роботи алгоритму Random Forest [20]

Якість роботи Random Forest значною мірою визначається значеннями його гіперпараметрів. До найважливіших із них належать кількість дерев в ансамблі, максимальна глибина дерев, мінімальна кількість об'єктів у вузлі та кількість ознак, що випадково відбираються під час кожного розбиття.

Правильний вибір цих параметрів дозволяє досягти балансу між точністю прогнозування, швидкістю навчання та здатністю моделі узагальнювати закономірності на нових даних.

Основними перевагами Random Forest є висока точність прогнозування, стійкість до перенавчання, можливість роботи з великою кількістю ознак різних типів та відносно невисока чутливість до викидів і шуму в даних. Саме завдяки цим властивостям алгоритм набув широкого поширення в задачах прогнозування вартості нерухомості та інших об'єктів, де необхідно враховувати складні взаємозв'язки між великою кількістю факторів.

Іншим важливим напрямом розвитку ансамблевих методів є градієнтний бустинг (Gradient Boosting). На відміну від Random Forest, де всі дерева будуються незалежно одне від одного, у градієнтному бустингу моделі створюються послідовно. Кожне нове дерево навчається з урахуванням помилок, допущених попереднім ансамблем, і спрямоване на їх виправлення. Завдяки такому підходу модель поступово підвищує точність прогнозування, концентруючись на найбільш складних для прогнозування спостереженнях.

Основною ідеєю градієнтного бустингу є послідовна мінімізація функції втрат. Після побудови кожного дерева визначається похибка прогнозування, а наступне дерево формується таким чином, щоб максимально її зменшити. У результаті остаточно модель складається з великої кількості відносно простих дерев, які спільно забезпечують високу якість прогнозування. Водночас послідовний характер навчання робить алгоритм більш чутливим до перенавчання, тому під час побудови моделей використовують механізми регуляризації, обмеження глибини дерев, зменшення швидкості навчання (learning rate) та ранню зупинку навчання (early stopping) [21].

Однією з найвідоміших реалізацій градієнтного бустингу є алгоритм XGBoost (Extreme Gradient Boosting). Він розширює класичний підхід за

рахунок використання регуляризації, оптимізованих алгоритмів побудови дерев, підтримки паралельних обчислень та ефективної роботи з пропущеними значеннями. Це забезпечує високу швидкість навчання, стійкість моделі та значне підвищення точності прогнозування порівняно з базовим алгоритмом градієнтного бустингу.

Принцип роботи алгоритму XGBoost наведено на рисунку 2.5. Як видно з рисунка, дерева будуються послідовно, при цьому кожне наступне дерево моделює залишкову похибку попереднього ансамблю. Остаточний прогноз формується шляхом підсумовування внесків усіх побудованих дерев.

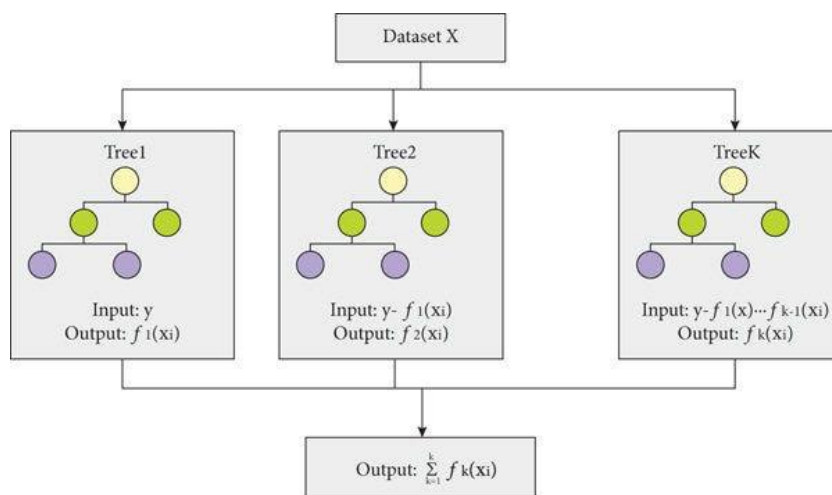


Рисунок 2.5 – Принцип роботи алгоритму XGBoost [22]

Ще одним сучасним алгоритмом сімейства градієнтного бустингу є CatBoost (Categorical Boosting), який був розроблений для ефективної роботи з категоріальними ознаками. На відміну від більшості інших алгоритмів, CatBoost не потребує попереднього кодування категоріальних змінних за допомогою методів One-Hot Encoding або Label Encoding. Замість цього алгоритм використовує власний механізм статистичного перетворення категоріальних ознак безпосередньо під час навчання моделі,

що дозволяє зменшити втрати інформації та підвищити якість прогнозування.

Крім ефективної обробки категоріальних ознак, CatBoost використовує спеціальну схему послідовного навчання та вдосконалений алгоритм побудови дерев, які дозволяють зменшити зміщення моделі та підвищити її узагальнювальну здатність. Завдяки цьому алгоритм демонструє високі результати на задачах регресії, особливо при роботі з наборами даних, що містять значну кількість категоріальних характеристик.

Загальну схему роботи алгоритму CatBoost наведено на рисунку 2.6. Вона ілюструє процес попередньої обробки категоріальних ознак, послідовного навчання дерев та формування остаточного прогнозу шляхом об'єднання результатів усіх побудованих моделей.

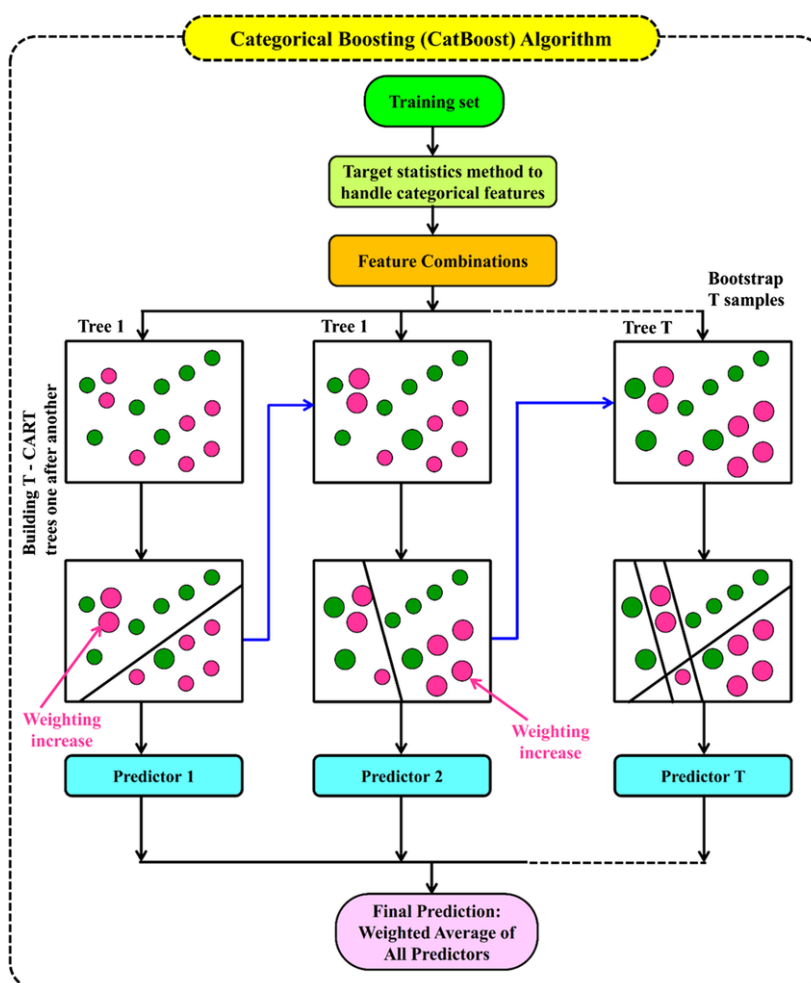


Рисунок 2.6 – Принцип роботи алгоритму CatBoost [22]

Сьогодні алгоритми XGBoost і CatBoost належать до найбільш ефективних методів машинного навчання для задач регресії. Вони широко застосовуються у фінансовому аналізі, прогнозуванні попиту, оцінюванні вартості нерухомості та інших прикладних задачах, де необхідно враховувати складні нелінійні залежності між великою кількістю ознак і забезпечувати високу точність прогнозування [22].

2.3 Метрики оцінювання якості регресійних моделей

Найбільш поширеними метриками оцінювання регресійних моделей є середньоквадратична похибка (MSE), середня абсолютна похибка (MAE), корінь із середньоквадратичної похибки (RMSE) та коефіцієнт детермінації (R^2) [23].

Середньоквадратична похибка (Mean Squared Error, MSE) визначає середнє значення квадратів різниці між фактичними та прогнозованими значеннями. Завдяки піднесенню похибок до квадрату ця метрика значно сильніше штрафує великі відхилення, ніж малі, тому широко використовується під час навчання регресійних моделей та оптимізації їх параметрів. Принцип визначення середньоквадратичної похибки наведено на рисунку 2.7.

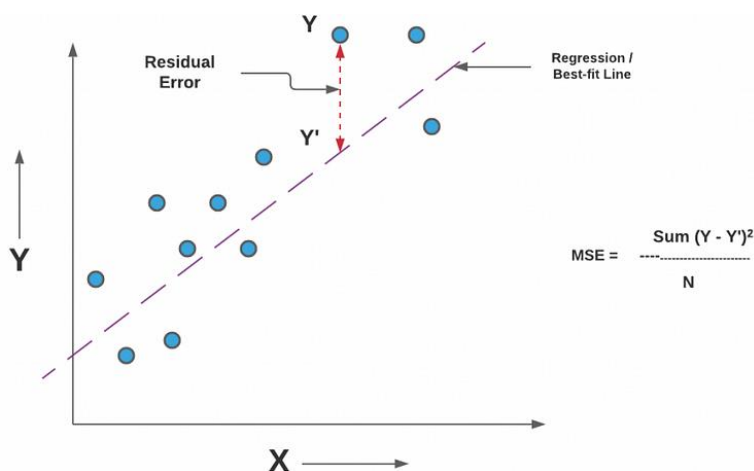


Рисунок 2.7 – Принцип визначення середньоквадратичної похибки [23]

Математично MSE визначається за формуло:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (2.2)$$

де n – кількість спостережень;

y_i – фактичне значення цільової змінної;

$\hat{y}_i$ – прогнозоване значення моделі.

Чим меншим є значення MSE, тим точніше модель описує досліджувані дані. Разом із тим через піднесення похибок до квадрату ця метрика є чутливою до викидів і поодиноких великих помилок прогнозування.

Середня абсолютна похибка (Mean Absolute Error, MAE) характеризує середнє абсолютне відхилення прогнозованих значень від фактичних без урахування напрямку похибки. На відміну від MSE, усі помилки мають однаковий внесок у кінцевий результат, тому MAE менш чутлива до великих відхилень.

Формула середньої абсолютної похибки має вигляд:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|, \quad (2.3)$$

де n – кількість спостережень;

y_i – фактичне значення;

$\hat{y}_i$ – прогнозоване значення моделі.

Основною перевагою MAE є проста інтерпретація, оскільки значення метрики виражається в тих самих одиницях вимірювання, що й цільова змінна. Водночас ця метрика менш чутлива до значних похибок, ніж MSE, тому великі відхилення не мають настільки суттєвого впливу на кінцевий результат.

Корінь із середньоквадратичної похибки (Root Mean Squared Error, RMSE) є модифікацією MSE та визначається як квадратний корінь із її значення. Завдяки цьому результат повертається до початкових одиниць вимірювання цільової змінної, що значно спрощує його практичну інтерпретацію.

Формула RMSE має вигляд:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n}}, \quad (2.4)$$

де n – кількість спостережень;

y_i – фактичне значення цільової змінної;

$\hat{y}_i$ – прогнозоване значення моделі.

Як і MSE, ця метрика сильніше реагує на великі похибки прогнозування, однак завдяки використанню початкових одиниць вимірювання є більш зрозумілою для практичного аналізу результатів.

Для оцінювання загальної якості регресійної моделі також широко використовується коефіцієнт детермінації (R^2). На відміну від попередніх метрик, він характеризує не абсолютну величину похибки, а частку варіації цільової змінної, яка пояснюється побудованою моделлю.

Коефіцієнт детермінації визначається за формулою:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (2.5)$$

де n – кількість спостережень;

y_i – фактичне значення цільової змінної;

$\hat{y}_i$ – прогнозоване значення моделі;

$\bar{y}$ – середнє значення цільової змінної.

Значення коефіцієнта детермінації зазвичай належить інтервалу від 0 до 1, де значення, близьке до 1, свідчить про високу здатність моделі

пояснювати зміну цільової змінної. Чим більшим є значення R^2 , тим краще модель описує наявні дані. Водночас використання лише цієї метрики не дозволяє оцінити абсолютну величину похибки прогнозування.

Таким чином, жодна окрема метрика не може повністю охарактеризувати якість регресійної моделі. MSE та RMSE ефективно виявляють значні похибки прогнозування, MAE характеризує середнє абсолютне відхилення без надмірного впливу викидів, а коефіцієнт детермінації R^2 оцінює здатність моделі пояснювати зміну цільової змінної. Саме тому для комплексного оцінювання якості моделей доцільно використовувати сукупність кількох метрик [23].

2.4 Методи інтерпретації результатів машинного навчання

Сучасні моделі машинного навчання забезпечують високу точність прогнозування, однак багато з них мають складну внутрішню структуру, що ускладнює пояснення процесу прийняття рішень. Такі моделі часто називають моделями типу «чорної скриньки» (black box), оскільки користувач отримує прогноз, але не може безпосередньо визначити, які саме ознаки та якою мірою вплинули на його формування. Для задач прогнозування вартості об'єктів це є суттєвим недоліком, оскільки кінцевий результат повинен бути не лише точним, а й зрозумілим для подальшого аналізу.

Для підвищення прозорості моделей використовується напрям Explainable Artificial Intelligence (XAI), який об'єднує методи інтерпретації результатів машинного навчання. Основною метою XAI є пояснення причин отримання певного прогнозу, визначення важливості окремих ознак та оцінювання їхнього впливу на кінцевий результат. Це дозволяє не лише підвищити довіру до моделі, а й виявити потенційні помилки під час побудови моделей та аналізу даних.

Одним із найбільш поширених сучасних підходів до інтерпретації моделей машинного навчання є метод SHAP (SHapley Additive exPlanations), запропонований на основі теорії кооперативних ігор. В основу методу покладено значення Шеплі (Shapley values), які використовуються для визначення справедливого внеску кожної ознаки у формування кінцевого прогнозу. Ідея полягає в тому, що кожна ознака розглядається як «учасник гри», а її внесок оцінюється шляхом аналізу зміни прогнозу при додаванні або видаленні цієї ознаки з моделі [24].

Для кожного прогнозу SHAP обчислює внесок усіх ознак окремо, показуючи не лише силу їхнього впливу, а й напрям цього впливу. Додатні значення SHAP свідчать про збільшення прогнозованого значення цільової змінної, тоді як від'ємні – про його зменшення. Завдяки цьому метод забезпечує як локальну інтерпретацію окремих прогнозів, так і глобальний аналіз поведінки моделі на всьому наборі даних.

Однією з основних переваг SHAP є можливість наочного представлення результатів інтерпретації за допомогою спеціалізованих візуалізацій. Найбільш поширеною серед них є SHAP Summary Plot, який одночасно демонструє важливість ознак, напрям їхнього впливу та розподіл внесків для всіх об'єктів дослідження. Приклад такої візуалізації наведено на рисунку 2.8.

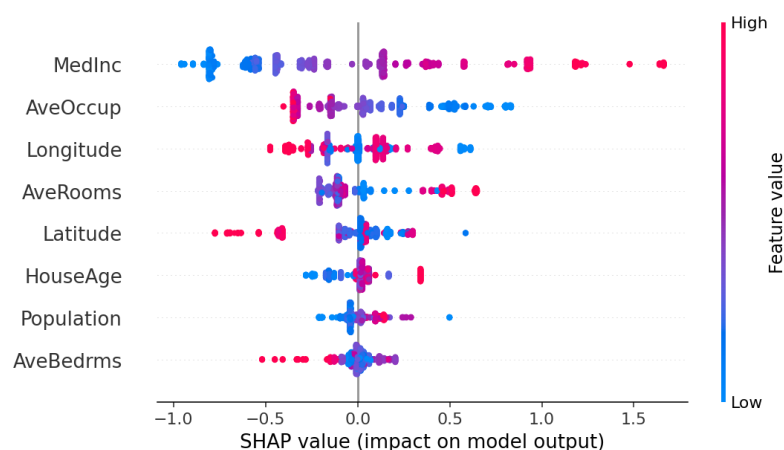


Рисунок 2.8 – Приклад візуалізації SHAP Summary Plot [24]

На рисунку 2.8 по вертикальній осі розташовано ознаки моделі в порядку зменшення їхньої важливості. Горизонтальна вісь відображає значення SHAP, яке характеризує величину та напрям впливу кожної ознаки на прогноз. Додатні значення збільшують прогнозовану вартість об'єкта, тоді як від'ємні – зменшують її. Кожна точка відповідає окремому спостереженню, а її колір відображає значення відповідної ознаки: синій колір відповідає низьким значенням, а червоний – високим. Такий спосіб подання результатів дозволяє одночасно оцінити важливість ознак, характер їхнього впливу та закономірності, що формуються під час роботи моделі.

Перевагами методу SHAP є універсальність, можливість інтерпретації практично будь-яких моделей машинного навчання, забезпечення як локального, так і глобального аналізу, а також наочне представлення отриманих результатів. Разом із тим для складних ансамблевих моделей обчислення значень SHAP може потребувати значних обчислювальних ресурсів, що є основним обмеженням цього підходу.

Завдяки поєднанню високої інформативності, універсальності та наочності SHAP сьогодні є одним із найпоширеніших методів інтерпретації моделей машинного навчання та широко використовується для аналізу результатів регресійних моделей, зокрема у задачах прогнозування вартості об'єктів.

3 РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ПРОГНОЗУВАННЯ ВАРТОСТІ ОБ'ЄКТІВ

У процесі розробки інтелектуальної системи прогнозування вартості об'єктів нерухомості використовувалися сучасні програмні засоби та бібліотеки машинного навчання, що забезпечують повний цикл аналізу даних – від їх підготовки до інтерпретації отриманих результатів.

Розробку програмної реалізації виконано мовою програмування Python, яка є одним із найбільш поширених інструментів для розв'язання задач аналізу даних та машинного навчання. Виконання програмного коду здійснювалося у хмарному середовищі Google Colab, що забезпечує доступ до необхідних обчислювальних ресурсів без встановлення додаткового програмного забезпечення на локальному комп'ютері.

Для роботи з табличними даними використовувалися бібліотеки Pandas та NumPy, які забезпечують ефективне завантаження, обробку та аналіз даних. Побудову графіків і візуалізацію результатів експериментів реалізовано за допомогою бібліотек Matplotlib та Seaborn.

Основні алгоритми машинного навчання реалізовано із використанням бібліотеки scikit-learn, яка застосовувалася для попередньої обробки даних, побудови конвеєра обробки ознак, навчання моделей Linear Regression, Random Forest та Extra Trees, а також для оцінювання якості прогнозування й оптимізації гіперпараметрів методом RandomizedSearchCV.

Для реалізації моделей градієнтного бустингу використовувалися спеціалізовані бібліотеки XGBoost та CatBoost, які містять високопродуктивні реалізації відповідних алгоритмів та широко застосовуються під час розв'язання задач регресії на табличних даних.

Інтерпретацію результатів прогнозування виконано за допомогою бібліотеки SHAP, що дозволяє визначити вплив окремих ознак на прогноз моделі та забезпечує пояснюваність отриманих результатів.

Використання зазначених програмних засобів дозволило реалізувати повний цикл експериментального дослідження, включаючи підготовку даних, навчання моделей машинного навчання, оцінювання їх ефективності, оптимізацію гіперпараметрів та аналіз факторів, що впливають на прогнозовану вартість об'єктів нерухомості.

3.1 Архітектура інтелектуальної системи прогнозування

Розроблена інтелектуальна система реалізована у вигляді послідовного ML pipeline, який об'єднує всі основні етапи прогнозування вартості об'єктів нерухомості. До складу системи входять модулі підготовки даних, формування ознак, навчання моделей машинного навчання, оцінювання їх ефективності, вибору фінальної моделі та інтерпретації отриманих результатів. Такий підхід забезпечує автоматизоване виконання повного циклу прогнозування та створює єдине середовище для проведення експериментів із різними алгоритмами машинного навчання.

Загальну архітектуру розробленої інтелектуальної системи прогнозування вартості об'єктів нерухомості наведено на рисунку 3.1.

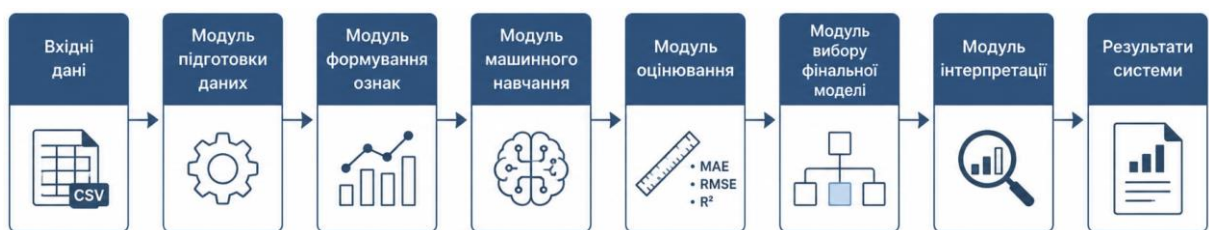


Рисунок 3.1 – Архітектура інтелектуальної системи прогнозування вартості об'єктів нерухомості

Як показано на рисунку 3.1, робота системи починається із завантаження вхідних даних у форматі CSV, після чого інформація передається до модуля підготовки даних. На цьому етапі здійснюється

аналіз структури набору даних, обробка пропущених значень, розподіл даних на навчальну та тестову вибірки, а також перетворення категоріальних ознак у числовий формат.

Після завершення попередньої обробки дані надходять до модуля формування ознак. На цьому етапі створюються додаткові характеристики об'єктів нерухомості, які дозволяють точніше описати їх властивості та підвищити інформативність навчальної вибірки. Отриманий набір ознак використовується для навчання моделей машинного навчання.

У модулі машинного навчання реалізовано навчання декількох моделей регресії, що дозволяє порівняти ефективність різних алгоритмів прогнозування вартості нерухомості. Після завершення навчання кожна модель передається до модуля оцінювання, у якому якість прогнозування визначається за метриками MAE, RMSE та коефіцієнтом детермінації R^2 .

На основі результатів оцінювання здійснюється вибір фінальної моделі, яка демонструє найкращі показники точності прогнозування. Обрана модель використовується для подальшого аналізу та формування остаточних прогнозів.

Заключним етапом роботи системи є модуль інтерпретації результатів прогнозування. На цьому етапі аналізується внесок окремих ознак у формування прогнозу, визначаються фактори, що найбільше впливають на вартість об'єктів нерухомості, та формуються засоби пояснення роботи моделі. Результатом функціонування системи є прогнозована вартість об'єкта нерухомості, показники точності моделі та візуальні матеріали, що відображають якість прогнозування й інтерпретацію отриманих результатів.

Запропонована архітектура забезпечує логічну послідовність виконання всіх етапів машинного навчання, дозволяє проводити комплексне експериментальне дослідження різних моделей прогнозування та створює основу для подальшого розширення функціональних можливостей інтелектуальної системи.

3.2 Формування експериментального набору даних

Для проведення дослідження використано набір даних Ames Housing Dataset [25], який містить інформацію про реальні житлові об'єкти та широко застосовується для дослідження методів прогнозування вартості нерухомості.

Завантаження набору даних здійснювалося з CSV-файлу за допомогою бібліотеки Pandas. Після імпорту було виконано первинну перевірку коректності завантаження, визначено розмір набору даних, перелік ознак та перевірено наявність цільової змінної SalePrice. У результаті встановлено, що експериментальний набір містить 2930 записів та 82 стовпці, серед яких один є цільовою ознакою, а решта описують характеристики об'єктів нерухомості.

Перші записи експериментального набору даних наведено на рисунку 3.2.

er	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	...	Pool Area	Pool QC	Fence	Misc Feature	Misc Val	Mo Sold	Yr Sold	Sale Type	Sale Condition	SalePrice
1	526301100	20	RL	141.0	31770	Pave	NaN	IR1	Lvl	...	0	NaN	NaN	NaN	0	5	2010	WD	Normal	215000
2	526350040	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	...	0	NaN	MnPrv	NaN	0	6	2010	WD	Normal	105000
3	526351010	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	...	0	NaN	NaN	Gar2	12500	6	2010	WD	Normal	172000
4	526353030	20	RL	93.0	11160	Pave	NaN	Reg	Lvl	...	0	NaN	NaN	NaN	0	4	2010	WD	Normal	244000
5	527105010	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	...	0	NaN	MnPrv	NaN	0	3	2010	WD	Normal	189900

Рисунок 3.2 – Перші записи експериментального набору даних Ames Housing Dataset

Наведені записи демонструють структуру набору даних та різноманітність представлених ознак. До складу вибірки входять як кількісні характеристики (площа земельної ділянки, площа будинку, кількість кімнат, рік побудови тощо), так і категоріальні параметри, що описують якість будівництва, тип покрівлі, матеріали оздоблення, район розташування та інші особливості нерухомості.

Після завантаження даних було виконано аналіз основних статистичних характеристик цільової змінної SalePrice. Отримані результати наведено на рисунку 3.3.

SalePrice	
count	2930.000000
mean	180796.060068
std	79886.692357
min	12789.000000
25%	129500.000000
50%	160000.000000
75%	213500.000000
max	755000.000000

Рисунок 3.3 – Основні статистичні характеристики цільової змінної SalePrice

З аналізу статистичних характеристик встановлено, що середня вартість об'єктів нерухомості становить 180796 доларів США, тоді як медіанне значення дорівнює 160000 доларів США. Мінімальна ціна продажу складає 12789 доларів США, а максимальна – 755000 доларів США. Значення стандартного відхилення 79886,69 свідчить про значну варіативність вартості об'єктів, що характерно для реального ринку нерухомості.

Для оцінювання особливостей розподілу цільової змінної було побудовано гістограму вартості об'єктів нерухомості.

Як видно з рисунка 3.4, розподіл вартості об'єктів нерухомості не є симетричним і характеризується правосторонньою асиметрією. Основна частина об'єктів має вартість у межах приблизно від 100 до 250 тисяч доларів США, тоді як дорожчі будинки представлені значно меншою кількістю спостережень. Такий характер розподілу є типовим для задач прогнозування вартості нерухомості та свідчить про наявність окремих дорогих об'єктів, які формують довгий правий хвіст розподілу.

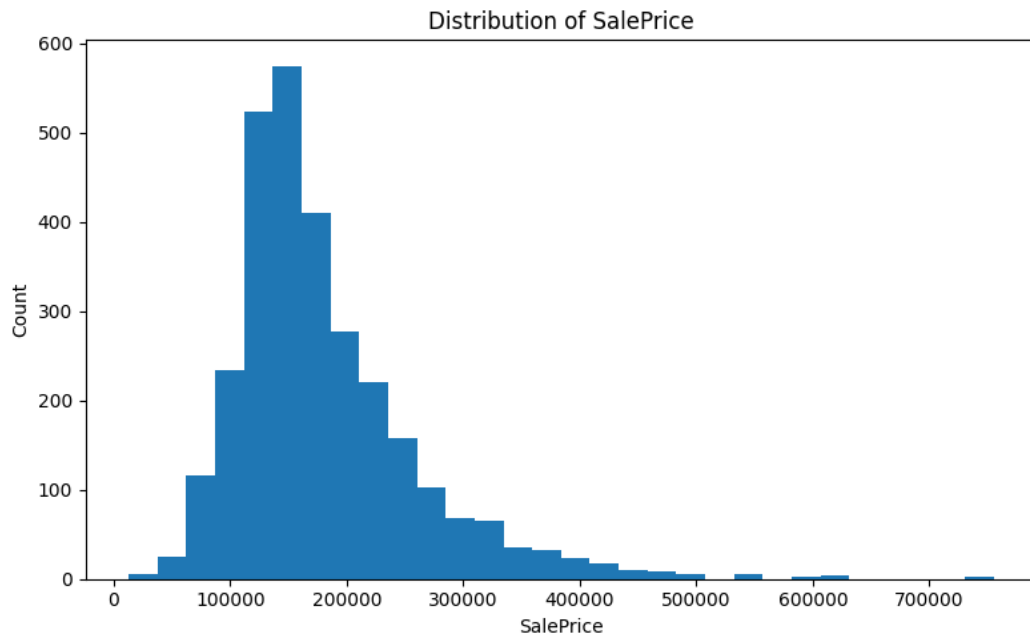


Рисунок 3.4 – Розподіл цільової змінної SalePrice

Наступним етапом було проведено аналіз якості експериментального набору даних. Встановлено, що набір містить 2930 записів та 82 ознаки, при цьому дублікати записів відсутні. Аналіз пропущених значень показав, що пропуски присутні у 27 ознаках, що обумовлює необхідність їх подальшої обробки на етапі попередньої підготовки даних. Крім того, виконано класифікацію ознак за типом даних. До початкового набору входять 39 числових та 43 категоріальні ознаки.

Після завершення первинного аналізу було сформовано набір ознак, який використовувався для навчання моделей машинного навчання. Із загального переліку стовпців виключено службові поля Order та PID, оскільки вони є ідентифікаторами записів і не містять інформації, що характеризує властивості об'єктів нерухомості. Також із матриці ознак вилучено цільову змінну SalePrice, значення якої використовувалося лише як прогнозована величина.

У результаті сформовано матрицю ознак розміром 2930×79 , що включає 36 числових та 43 категоріальні ознаки, а також окремий вектор цільових значень розміром 2930×1 .

Для оцінювання здатності моделей узагальнювати закономірності на нових даних сформований набір було розділено на навчальну та тестову вибірки із використанням функції `train_test_split` бібліотеки `scikit-learn`. Навчальна вибірка містить 2344 спостереження (80%), тоді як тестова – 586 спостережень (20%). Такий розподіл є одним із найбільш поширених у задачах машинного навчання та забезпечує достатній обсяг даних як для навчання моделей, так і для їх незалежного оцінювання.

Результати розподілу експериментального набору наведено на рисунку 3.5.

Train:		Test:	
count	2344.000000	count	586.000000
mean	178582.207765	mean	189651.469283
std	77125.072713	std	89617.259169
min	12789.000000	min	44000.000000
25%	129000.000000	25%	130000.000000
50%	160000.000000	50%	165000.000000
75%	210000.000000	75%	228375.000000
max	755000.000000	max	625000.000000

Рисунок 3.5 – Результати розподілу експериментального набору на навчальну та тестову вибірки

Порівняння статистичних характеристик цільової змінної для навчальної та тестової вибірок показало, що їх основні показники є близькими між собою. Це свідчить про збереження загального характеру розподілу вартості нерухомості після розділення набору даних і підтверджує коректність формування навчальної та тестової вибірок для проведення подальших експериментальних досліджень.

3.3 Попередня обробка даних та формування ознак

Попередня обробка даних є необхідним етапом побудови моделей машинного навчання, оскільки більшість алгоритмів не можуть безпосередньо працювати з пропущеними значеннями та категоріальними

ознаками. Крім того, якість підготовки даних істотно впливає на точність прогнозування, тому для забезпечення єдиного підходу до їх обробки було реалізовано автоматизований конвеєр попередньої підготовки.

Для реалізації конвеєра використано механізм Pipeline бібліотеки scikit-learn, який забезпечує послідовне виконання всіх етапів перетворення даних. Оскільки набір Ames Housing Dataset містить як числові, так і категоріальні ознаки, їх обробку організовано за допомогою ColumnTransformer, що дозволяє застосовувати різні методи попередньої обробки до різних типів даних.

На першому етапі реалізовано базовий конвеєр попередньої обробки. Для числових ознак пропущені значення замінювалися медіанними значеннями відповідних стовпців із використанням SimpleImputer зі стратегією median. Такий підхід є стійким до наявності аномальних значень та не призводить до істотного зміщення розподілу даних.

Для категоріальних ознак застосовано двоетапну обробку. Спочатку пропущені значення замінювалися найбільш поширеним значенням відповідної ознаки (most frequent), після чого виконувалося кодування методом One-Hot Encoding. Використання параметра handle_unknown="ignore" забезпечує коректну роботу системи у випадку появи нових категорій, які були відсутні під час навчання моделі.

Після побудови базового конвеєра було виконано його перевірку на навчальній та тестовій вибірках. Результати наведено на рисунку 3.6.

```
Original X_train shape: (2344, 79)
Processed X_train shape: (2344, 300)
Original X_test shape: (586, 79)
Processed X_test shape: (586, 300)

Missing values after preprocessing:
Train: 0
Test : 0
```

Рисунок 3.6 – Результат застосування базового конвеєра попередньої обробки даних

Як видно з рисунка 3.6, після виконання One-Hot Encoding кількість ознак збільшилася з 79 до 300, що обумовлено перетворенням кожного можливого значення категоріальних ознак на окрему бінарну змінну. При цьому кількість записів у навчальній та тестовій вибірках залишилася незмінною. Додаткова перевірка підтвердила повну відсутність пропущених значень після завершення попередньої обробки, що забезпечує можливість подальшого навчання моделей машинного навчання.

Після реалізації базового підходу було розроблено вдосконалений конвеєр попередньої обробки, який враховує особливості окремих категоріальних ознак набору Ames Housing Dataset. Аналіз предметної галузі показав, що для низки характеристик відсутність значення не означає втрату інформації, а свідчить про фактичну відсутність відповідного конструктивного елемента будинку. До таких ознак належать, зокрема, наявність провулку (Alley), характеристики підвалу (Bsmt Qual, Bsmt Cond, Bsmt Exposure, BsmtFin Type 1, BsmtFin Type 2), тип та стан гаража (Garage Type, Garage Finish, Garage Qual, Garage Cond), якість басейну (Pool QC), тип огорожі (Fence), додаткові споруди (Misc Feature) та інші.

Для цих п'ятнадцяти ознак пропущені значення замінювалися спеціальним категоріальним значенням None, що дозволяє моделі відрізняти відсутність конструктивного елемента від невідомого значення. Для решти категоріальних ознак, де пропуски є наслідком неповноти даних, використовувалася стратегія заповнення найбільш поширеним значенням. Такий підхід дозволяє зберегти семантичний зміст даних і забезпечує більш коректне навчання моделей.

Наступним етапом стало формування нових ознак (feature engineering), спрямоване на підвищення інформативності початкового набору даних. Замість використання лише первинних характеристик було сформовано додаткові узагальнені показники, які більш повно описують властивості житлових будинків.

Для врахування віку будинку створено ознаку House Age, яка визначається як різниця між роком продажу та роком побудови будинку. Аналогічно сформовано ознаку Remod Age, що характеризує кількість років, які минули після останньої реконструкції об'єкта до моменту його продажу. Такі показники дозволяють врахувати фізичний знос будівлі та актуальність виконаних ремонтних робіт.

Для комплексної оцінки розміру житлового простору сформовано ознаку Total SF, яка об'єднує площу підвалу, першого та другого поверхів. На відміну від окремих площ поверхів, цей показник характеризує загальний корисний простір будинку та є більш інформативним для прогнозування його вартості.

Додатково сформовано ознаку Total Porch SF, яка визначає сумарну площу всіх типів веранд і терас незалежно від їх конструктивного виконання. Такий показник дозволяє оцінити загальну площу відкритих прибудов, що також може впливати на ринкову вартість нерухомості.

Для більш повної характеристики санітарних приміщень створено ознаку Total Bathrooms. Під час її обчислення повноцінні ванні кімнати враховувалися з коефіцієнтом 1, тоді як напівванні кімнати – з коефіцієнтом 0,5. При цьому враховувалися як основні, так і підвальні санітарні вузли. Такий підхід забезпечує більш реалістичне відображення функціональних можливостей будинку.

Крім кількісних характеристик було сформовано чотири бінарні ознаки: Has Garage, Has Basement, Has Fireplace та Has Pool. Кожна з них набуває значення 1 за наявності відповідного конструктивного елемента та 0 у протилежному випадку. Використання таких ознак дозволяє моделі окремо враховувати сам факт наявності гаража, підвалу, каміна або басейну незалежно від їх кількісних характеристик.

Останньою сформованою ознакою стала SF per Room, яка визначає середню площу житлового приміщення як відношення житлової площі до кількості кімнат. Цей показник характеризує просторість будинку та

дозволяє відрізнити будинки з однаковою загальною площею, але різним плануванням. Для уникнення ділення на нуль відсутні значення автоматично замінювалися медіанним значенням сформованої ознаки.

У результаті виконання feature engineering кількість ознак збільшилася з 79 до 89, тобто було сформовано 10 додаткових характеристик. При цьому кількість числових ознак зросла з 36 до 46, тоді як кількість категоріальних ознак залишилася незмінною і становила 43. Результати формування нових ознак наведено на рисунку 3.7.

```
Original X_train shape: (2344, 79)
Engineered X_train shape: (2344, 89)
Original X_test shape: (586, 79)
Engineered X_test shape: (586, 89)
Added features: 10

Numeric features after FE: 46
Categorical features after FE: 43

New engineered columns:
- House Age
- Remod Age
- Total SF
- Total Porch SF
- Total Bathrooms
- Has Garage
- Has Basement
- Has Fireplace
- Has Pool
- SF per Room
```

Рисунок 3.7 – Результати формування нових ознак після виконання feature engineering

Як видно з рисунка 3.7, застосування feature engineering дозволило збільшити кількість ознак з 79 до 89 без зміни кількості об'єктів у навчальній та тестовій вибірках. Це забезпечило розширення інформаційного опису об'єктів нерухомості та створило передумови для підвищення точності моделей машинного навчання.

Після формування нових ознак було побудовано оновлений конвеєр попередньої обробки, який використовував розширений набір числових характеристик і той самий підхід до обробки категоріальних

даних. Отриманий конвеєр використовувався під час навчання всіх моделей машинного навчання в експериментальному дослідженні.

3.4 Реалізація моделей машинного навчання та експериментального дослідження

Після підготовки експериментального набору даних та реалізації конвеєрів попередньої обробки було виконано програмну реалізацію моделей машинного навчання для прогнозування вартості об'єктів нерухомості. Основною вимогою до реалізації було забезпечення однакових умов навчання, оцінювання та порівняння моделей. Для цього всі моделі інтегрувалися у єдиний конвеєр машинного навчання, який об'єднував етап попередньої обробки даних і безпосередньо регресійну модель.

Для оцінювання якості прогнозування було реалізовано окрему службову функцію, яка обчислює три метрики регресії: середню абсолютну похибку MAE, корінь із середньоквадратичної похибки RMSE та коефіцієнт детермінації R^2 . Використання єдиної функції для обчислення метрик дозволило уникнути дублювання коду та забезпечило однаковий підхід до оцінювання всіх моделей.

Окрім обчислення числових метрик, у системі реалізовано універсальну функцію оцінювання `evaluate_pipeline()`. Вона виконує повний цикл роботи з моделлю: навчання конвеєра на навчальній вибірці, формування прогнозів для тестової вибірки, обчислення метрик якості, а також повернення результатів у стандартизованому форматі. За потреби ця функція також забезпечує збереження навченої моделі у файл за допомогою бібліотеки `joblib`.

Для побудови моделей використано функцію `build_pipeline()`, яка формує єдиний об'єкт `Pipeline` із двох основних етапів: попередньої обробки даних та регресійної моделі. Завдяки цьому кожна модель навчалася не на вручну підготовлених масивах, а в межах повного конвеєра, що є більш

коректним підходом для експериментального дослідження. Така структура зменшує ризик витоку даних між навчальною та тестовою вибірками, оскільки всі перетворення виконуються всередині pipeline.

У програмній реалізації було зафіксовано назви основних моделей, що використовувалися в експериментах: Linear Regression, Random Forest, Extra Trees, XGBoost та CatBoost. Також створено спільне сховище результатів експериментів, у якому накопичувалися значення метрик для кожної моделі та кожного етапу дослідження.

Для збереження результатів було реалізовано функцію `save_experiment_result()`. Вона додає результат чергового експерименту до загальної структури, формує підсумкову таблицю, сортує її за назвою експерименту та моделлю, а також зберігає отримані дані у форматі CSV. Це забезпечило централізоване накопичення результатів і спростило подальший порівняльний аналіз моделей у четвертому розділі.

Першою реалізованою моделлю стала Linear Regression, яка використовувалася як базова модель для порівняння. Її застосування дозволило отримати початковий рівень якості прогнозування та надалі оцінювати, наскільки складніші ансамблеві моделі покращують результат. Лінійну регресію було послідовно використано у трьох варіантах: із базовим конвеєром попередньої обробки, з покращеним preprocessing та з розширеним набором ознак після feature engineering.

Далі було реалізовано модель Random Forest Regressor, яка належить до ансамблевих методів на основі дерев рішень. У моделі було задано 300 дерев, фіксоване значення `random_state` для відтворюваності результатів та параметр `n_jobs=-1` для використання доступних обчислювальних ресурсів. Random Forest було застосовано як до даних із покращеним preprocessing, так і до даних після формування нових ознак, що дозволило дослідити вплив feature engineering на якість ансамблевої моделі.

Наступною реалізованою моделлю стала Extra Trees Regressor. Як і Random Forest, ця модель є ансамблем дерев, однак використовує вищий

рівень випадковості під час побудови дерев. Для неї також було задано 300 дерев, фіксований `random_state` та паралельне виконання обчислень. Extra Trees використовувалася на даних після `feature engineering`, тобто на розширеному наборі ознак.

Для реалізації градієнтного бустингу використано модель XGBoost Regressor. У початковому варіанті для неї було задано 300 дерев, швидкість навчання 0,05, максимальну глибину дерева 6, цільову функцію `reg:squarederror`, фіксований `random_state` та паралельне виконання обчислень. Модель XGBoost також навчалася на даних після формування нових ознак.

Окремо було реалізовано модель CatBoost Regressor, яка також належить до методів градієнтного бустингу. Для початкового запуску було використано 300 ітерацій, швидкість навчання 0,05, глибину дерева 6 та фіксоване значення `random_state`. Параметр `verbose=0` було задано для вимкнення проміжного виведення під час навчання моделі.

Після реалізації базових варіантів моделей було проведено етап налаштування гіперпараметрів для XGBoost та CatBoost. Для цього застосовано метод `RandomizedSearchCV`, який виконує випадковий перебір комбінацій параметрів із заданого простору пошуку та оцінює їх за допомогою крос-валідації. Як цільову метрику для оптимізації було використано від'ємне значення RMSE (`neg_root_mean_squared_error`), що дозволяє шукати комбінацію параметрів, яка мінімізує похибку прогнозування.

Для XGBoost було сформовано простір пошуку, який включав кількість дерев, максимальну глибину, швидкість навчання, частку об'єктів для навчання окремих дерев, частку ознак для побудови дерева та мінімальну вагу дочірнього вузла. Пошук виконувався для 20 випадкових комбінацій параметрів із використанням 5-кратної крос-валідації.

Для CatBoost також було реалізовано налаштування гіперпараметрів за допомогою `RandomizedSearchCV`. Простір пошуку включав кількість

ітерацій, глибину дерева, швидкість навчання, коефіцієнт L2-регуляризації, параметр bagging temperature та параметр random strength. Як і для XGBoost, було використано 20 випадкових комбінацій параметрів і 5-кратну крос-валідацію.

Після завершення пошуку найкращі знайдені конфігурації XGBoost та CatBoost повторно оцінювалися на тестовій вибірці за єдиною процедурою `evaluate_pipeline()`. Для фінальної моделі CatBoost додатково було увімкнено збереження навченого `pipeline` у файл, що дозволяє використовувати модель повторно без необхідності повторного навчання.

Таким чином, у межах програмної реалізації було створено єдину систему навчання та оцінювання моделей машинного навчання. Усі моделі навчалися в межах стандартизованого `pipeline`, оцінювалися за однаковими метриками та зберігали результати в єдиному форматі. Це забезпечило коректність подальшого порівняльного аналізу та дозволило об'єктивно визначити модель, яка найкраще розв'язує задачу прогнозування вартості об'єктів нерухомості.

4 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ

Після реалізації інтелектуальної системи було проведено експериментальне дослідження, метою якого є оцінювання ефективності запропонованих підходів до прогнозування вартості об'єктів нерухомості. Послідовність експериментів відповідає етапам розробки системи та передбачає дослідження впливу попередньої обробки даних, формування нових ознак, вибору моделі машинного навчання, налаштування гіперпараметрів і аналіз інтерпретованості отриманих прогнозів.

Порівняння моделей виконувалося за єдиною методикою із використанням однакових навчальної та тестової вибірок і спільного набору метрик якості. Такий підхід забезпечує коректне зіставлення отриманих результатів та дозволяє об'єктивно оцінити вплив кожного етапу експериментального дослідження на точність прогнозування.

4.1 Дослідження впливу попередньої обробки даних та формування ознак

Перший етап експериментального дослідження був спрямований на оцінювання впливу попередньої обробки даних на якість прогнозування. Для забезпечення коректності порівняння в обох експериментах використовувалася одна й та сама модель Linear Regression, тоді як відмінність полягала виключно у способі підготовки даних.

У першому експерименті застосовано базовий конвеєр попередньої обробки, який передбачав заповнення пропущених числових значень медіаною, категоріальних – найчастішим значенням та подальше кодування категоріальних ознак методом One-Hot Encoding. У другому експерименті використано вдосконалений конвеєр, у якому для п'ятнадцяти

категоріальних ознак пропущені значення інтерпретувалися як відсутність відповідного конструктивного елемента та замінювалися категорією None.

Результати експерименту наведено на рисунку 4.1.

	Variant	Model	MAE	RMSE	R2
	Basic preprocessing	Linear Regression	16007.378341	29615.375198	0.890606
	Improved preprocessing	Linear Regression	15665.731611	29105.658651	0.894339

Рисунок 4.1 – Вплив попередньої обробки даних на якість прогнозування моделі Linear Regression

Як видно з рисунка 4.1, використання вдосконаленого підходу до попередньої обробки забезпечило покращення всіх досліджуваних показників. Значення середньої абсолютної похибки зменшилося з 16007,38 до 15665,73, а значення RMSE – з 29615,38 до 29105,66. Одночасно коефіцієнт детермінації збільшився з 0,8906 до 0,8943, що свідчить про незначне, але стабільне підвищення точності прогнозування.

Візуальне порівняння фактичних і прогнозованих значень для моделі після застосування вдосконаленої попередньої обробки наведено на рисунку 4.2.

Переважає більшість прогнозів розташована поблизу лінії ідеального прогнозування, що свідчить про достатньо високу узгодженість між фактичними та прогнозованими значеннями вартості об'єктів. Найбільші відхилення спостерігаються для окремих об'єктів із високою вартістю, що характерно для лінійних моделей при прогнозуванні нерухомості.

За результатами першого експерименту встановлено, що врахування семантики пропущених значень під час попередньої обробки даних дозволяє підвищити точність моделі навіть без зміни алгоритму машинного навчання. Отримані результати підтвердили доцільність використання вдосконаленого конвеєра попередньої обробки в подальших експериментах.

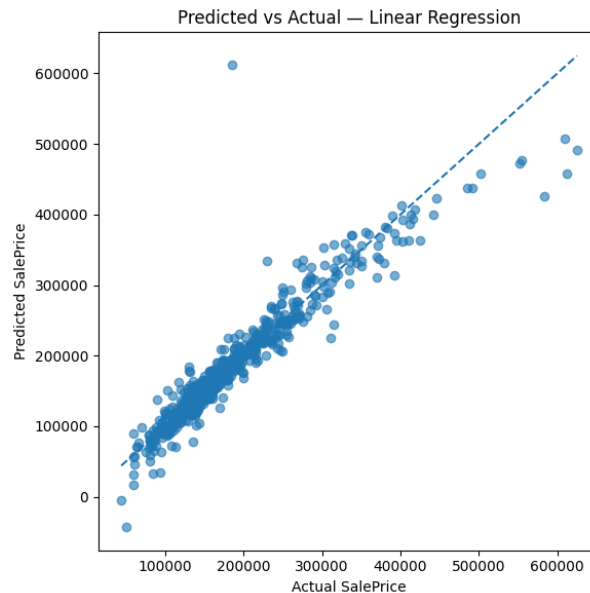


Рисунок 4.2 – Порівняння фактичних і прогнозованих значень для моделі Linear Regression після покращеної попередньої обробки даних

Після оцінювання впливу попередньої обробки даних було проведено другий експеримент, метою якого стало дослідження ефективності сформованих додаткових ознак (feature engineering). Для цього порівнювалися результати моделей Linear Regression та Random Forest до і після додавання нових ознак. В усіх інших аспектах умови проведення експерименту залишалися незмінними.

Результати експериментального дослідження наведено на рисунку 4.3.

Model	Variant	MAE	RMSE	R2
Linear Regression	Without feature engineering	15665.731611	29105.658651	0.894339
Linear Regression	With feature engineering	15807.406005	29940.701787	0.888190
Random Forest	Without feature engineering	15687.370722	26568.474361	0.911958
Random Forest	With feature engineering	15463.071149	25756.814226	0.917255

Рисунок 4.3 – Вплив feature engineering на якість прогнозування моделей машинного навчання

Як видно з рисунка 4.3, сформовані додаткові ознаки по-різному вплинули на якість прогнозування залежно від використаної моделі машинного навчання.

Для моделі Linear Regression застосування feature engineering не забезпечило покращення результатів. Значення середньої абсолютної похибки збільшилося з 15665,73 до 15807,41, значення RMSE – з 29105,66 до 29940,70, а коефіцієнт детермінації зменшився з 0,8943 до 0,8882. Такий результат можна пояснити тим, що більшість нових ознак була сформована на основі вже існуючих характеристик шляхом їх агрегування або обчислення похідних показників. Унаслідок цього між окремими ознаками виникла надлишковість інформації та посилювалися лінійні залежності, до яких модель лінійної регресії є чутливою. Як наслідок, сформовані ознаки не забезпечили приросту точності прогнозування.

Для моделі Random Forest спостерігалася протилежна тенденція. Після додавання нових ознак значення MAE зменшилося з 15687,37 до 15463,07, RMSE – з 26568,47 до 25756,81, а коефіцієнт детермінації збільшився з 0,9120 до 0,9173. Отримані результати свідчать, що ансамбль дерев рішень зміг ефективно використати сформовані агреговані характеристики. На відміну від лінійної регресії, дерева рішень добре працюють із похідними та узагальненими ознаками, автоматично враховуючи складні нелінійні взаємозв'язки між характеристиками об'єктів нерухомості.

Таким чином, результати другого експерименту підтвердили, що ефективність feature engineering залежить від типу використовуваної моделі машинного навчання. Для лінійної регресії сформовані додаткові ознаки не покращили якість прогнозування, тоді як для Random Forest вони виявилися корисними та забезпечили підвищення точності моделі.

З огляду на отримані результати, у подальших експериментах для ансамблевих моделей машинного навчання використовувався набір даних із

додатково сформованими ознаками, оскільки саме він забезпечив найкращі показники якості прогнозування.

4.2 Порівняльний аналіз моделей машинного навчання

Після дослідження впливу попередньої обробки даних та формування нових ознак було проведено порівняльний аналіз моделей машинного навчання з метою визначення алгоритму, який забезпечує найвищу точність прогнозування вартості об'єктів нерухомості. Для забезпечення коректності порівняння всі моделі навчалися на однаковому наборі даних із використанням покращеного preprocessing та сформованих додаткових ознак. Якість прогнозування оцінювалася за метриками MAE, RMSE та R^2 .

Результати експериментального дослідження наведено на рисунку 4.4.

	Model	MAE	RMSE	R2
0	CatBoost	13720.752850	22659.470328	0.935959
1	XGBoost	14211.511719	24149.194935	0.927262
2	Random Forest	15463.071149	25756.814226	0.917255
3	Extra Trees	15108.164124	27080.607477	0.908531
4	Linear Regression	15807.406005	29940.701787	0.888190

Рисунок 4.4 – Порівняльні результати моделей машинного навчання

Як видно з рисунка 4.4, найкращі результати серед усіх досліджуваних моделей продемонструвала CatBoost. Вона забезпечила найменші значення MAE (13720,75) та RMSE (22659,47), а також найбільше значення коефіцієнта детермінації ($R^2 = 0,9360$). Це свідчить про найвищу точність прогнозування та найкращу здатність моделі описувати залежність між характеристиками об'єктів нерухомості та їх ринковою вартістю.

Друге місце за якістю прогнозування посіла модель XGBoost, яка отримала значення MAE = 14211,51, RMSE = 24149,19 та $R^2 = 0,9273$. Хоча

її результати дещо поступаються CatBoost, модель також продемонструвала високу ефективність і значно перевершила класичні ансамблеві методи.

Модель Random Forest забезпечила значення RMSE 25756,81 та $R^2 = 0,9173$, що відповідає третьому місцю серед досліджуваних алгоритмів. Незважаючи на нижчу точність порівняно з моделями градієнтного бустингу, Random Forest продемонструвала стабільні результати та підтвердила ефективність ансамблевого підходу для задач прогнозування вартості нерухомості.

Модель Extra Trees показала дещо нижчу якість прогнозування порівняно з Random Forest. Значення RMSE становило 27080,61, а коефіцієнт детермінації – 0,9085. Це свідчить про те, що додаткова випадковість під час побудови дерев у цьому випадку не забезпечила покращення якості прогнозування.

Найнижчу точність серед досліджуваних моделей продемонструвала Linear Regression, для якої значення RMSE становило 29940,70, а коефіцієнт детермінації – 0,8882. Отриманий результат є очікуваним, оскільки лінійна регресія може описувати лише лінійні залежності між ознаками, тоді як ринок нерухомості характеризується складними нелінійними взаємозв'язками між численними факторами.

Для наочнішого порівняння моделей за значенням RMSE побудовано діаграму, наведену на рисунку 4.5.

Діаграма підтверджує результати табличного аналізу та демонструє поступове зменшення похибки прогнозування при переході від лінійної регресії до ансамблевих методів і моделей градієнтного бустингу. Найменше значення RMSE отримано для CatBoost, тоді як найбільше – для Linear Regression. При цьому різниця між XGBoost і CatBoost є меншою, ніж між лінійною регресією та ансамблевими моделями, що свідчить про загалом високу ефективність сучасних методів градієнтного бустингу для задач прогнозування вартості нерухомості.

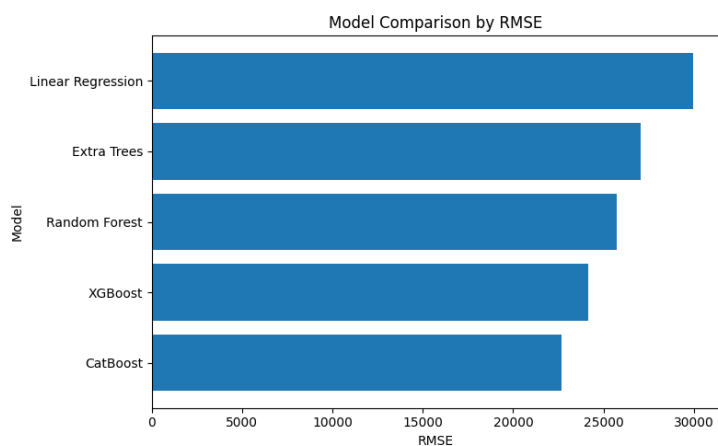


Рисунок 4.5 – Порівняння моделей машинного навчання за значенням RMSE

Візуальне порівняння фактичних і прогнозованих значень для всіх досліджуваних моделей наведено на рисунку 4.6.

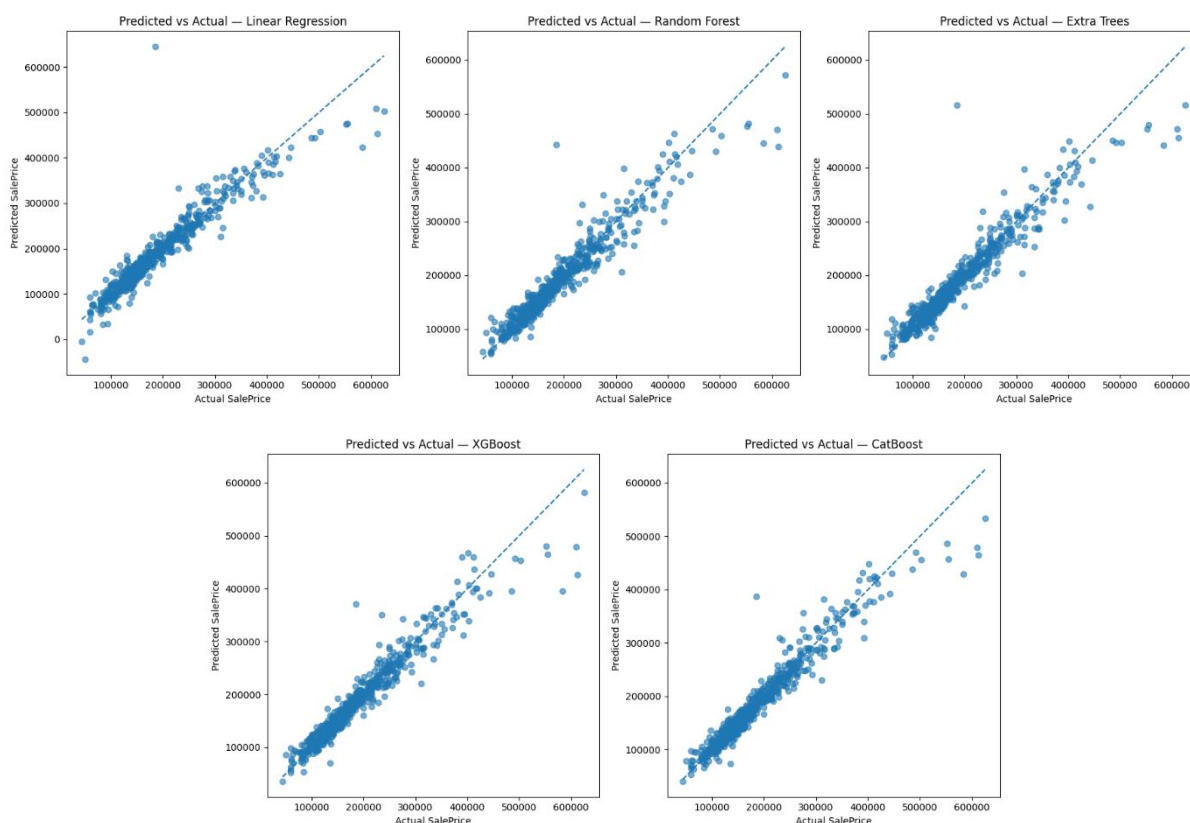


Рисунок 4.6 – Порівняння фактичних і прогнозованих значень для моделей машинного навчання

Як видно з рисунка 4.6, для моделі Linear Regression спостерігається найбільший розкид прогнозованих значень відносно лінії ідеального прогнозування. Особливо помітними є відхилення для об'єктів із високою вартістю, де модель схильна недооцінювати реальну ціну. Це пояснюється тим, що лінійна регресія не здатна повною мірою описати складні нелінійні залежності між характеристиками нерухомості.

Моделі Random Forest та Extra Trees демонструють більш щільне розташування точок уздовж діагоналі, що свідчить про суттєве підвищення точності прогнозування порівняно з лінійною регресією. При цьому для Random Forest відповідність між фактичними та прогнозованими значеннями є дещо кращою, що узгоджується з нижчим значенням RMSE та вищим коефіцієнтом детермінації.

Найменший розкид прогнозованих значень спостерігається для моделей XGBoost та CatBoost. Більшість точок розташована безпосередньо поблизу лінії ідеального прогнозування, що свідчить про високу точність побудованих моделей. Водночас для CatBoost концентрація точок біля діагоналі є найбільшою, особливо в діапазоні середніх значень вартості, що повністю відповідає отриманим кількісним результатам.

Разом із тим для всіх моделей зберігаються окремі значні відхилення, переважно для найдорожчих об'єктів нерухомості. Це свідчить про те, що прогнозування вартості об'єктів із крайніми значеннями залишається найбільш складною задачею незалежно від використаного алгоритму, хоча моделі градієнтного бустингу демонструють найкращу здатність працювати і з цією категорією об'єктів.

Таким чином, проведене порівняльне дослідження показало перевагу моделей градієнтного бустингу над іншими досліджуваними алгоритмами. Найкращі результати серед базових моделей продемонструвала CatBoost, однак різниця між нею та XGBoost виявилася незначною. З огляду на це обидві моделі були відібрані для подальшого налаштування гіперпараметрів та детальнішого експериментального дослідження.

4.3 Дослідження впливу налаштування гіперпараметрів

Після порівняльного аналізу моделей машинного навчання для подальшого дослідження було відібрано дві моделі градієнтного бустингу – XGBoost та CatBoost. Саме ці моделі продемонстрували найкращі результати на попередньому етапі експериментального дослідження, тому для них було виконано налаштування гіперпараметрів за допомогою методу RandomizedSearchCV.

Підбір гіперпараметрів здійснювався з використанням 5-кратної крос-валідації. Для кожної моделі перевірялося 20 випадкових комбінацій параметрів із заданого простору пошуку. Як критерій оптимізації використовувалося значення RMSE, оскільки ця метрика сильніше штрафує великі похибки прогнозування та є показовою для задачі оцінювання вартості нерухомості.

Результати пошуку найкращих параметрів для моделі XGBoost наведено на рисунку 4.7.

```
Best parameters:
model__subsample: 0.7
model__n_estimators: 300
model__min_child_weight: 3
model__max_depth: 4
model__learning_rate: 0.05
model__colsample_bytree: 0.7

Best CV RMSE: 22823.6977
```

Рисунок 4.7 – Результати підбору гіперпараметрів моделі XGBoost

За результатами RandomizedSearchCV для XGBoost найкращою виявилася конфігурація з кількістю дерев 300, максимальною глибиною дерева 4, швидкістю навчання 0,05, параметром subsample = 0,7,

colsample_bytree = 0,7 та min_child_weight = 3. Найкраще значення RMSE на крос-валідації становило 22823,70.

Результати пошуку найкращих параметрів для моделі CatBoost наведено на рисунку 4.8.

```
Best parameters:
model__random_strength: 1
model__learning_rate: 0.1
model__l2_leaf_reg: 9
model__iterations: 500
model__depth: 6
model__bagging_temperature: 3

Best CV RMSE: 22958.4435
```

Рисунок 4.8 – Результати підбору гіперпараметрів моделі CatBoost

Для CatBoost найкращими параметрами стали: кількість ітерацій 500, глибина дерева 6, швидкість навчання 0,1, коефіцієнт L2-регуляризації 9, bagging_temperature = 3 та random_strength = 1. Найкраще значення RMSE на крос-валідації становило 22958,44.

Після визначення найкращих конфігурацій обидві моделі було повторно навчено та оцінено на тестовій вибірці. Порівняння результатів до та після налаштування гіперпараметрів наведено на рисунку 4.9.

	Model	Variant	MAE	RMSE	R2
0	XGBoost	Before tuning	14211.511719	24149.194935	0.927262
1	XGBoost	After tuning	13490.539062	22825.134216	0.935019
2	CatBoost	Before tuning	13720.752850	22659.470328	0.935959
3	CatBoost	After tuning	13290.486515	22213.756197	0.938454

Рисунок 4.9 – Вплив налаштування гіперпараметрів на якість моделей XGBoost і CatBoost

Як видно з рисунка 4.9, налаштування гіперпараметрів покращило якість прогнозування для обох моделей. Для XGBoost значення MAE зменшилося з 14211,51 до 13490,54, RMSE – з 24149,19 до 22825,13, а коефіцієнт детермінації зріс із 0,9273 до 0,9350. Отже, tuning дозволив суттєво наблизити якість XGBoost до результатів CatBoost.

Для CatBoost також спостерігається покращення всіх метрик. Значення MAE зменшилося з 13720,75 до 13290,49, RMSE – з 22659,47 до 22213,76, а коефіцієнт детермінації збільшився з 0,9360 до 0,9385. Хоча приріст для CatBoost був менш різким, ніж для XGBoost, саме ця модель після налаштування показала найкращий абсолютний результат серед усіх досліджуваних моделей.

Порівняння результатів після налаштування показало, що CatBoost має найменші значення MAE та RMSE і найбільше значення R^2 . Тому саме налаштовану модель CatBoost було обрано як фінальну модель інтелектуальної системи прогнозування вартості об'єктів нерухомості. Подальший аналіз зосереджено на інтерпретації її прогнозів та дослідженні характеру похибок.

4.4 Інтерпретація результатів прогнозування

Після проведення порівняльного аналізу моделей та налаштування гіперпараметрів як фінальну модель інтелектуальної системи було обрано CatBoost, яка продемонструвала найкращі значення всіх метрик якості прогнозування ($MAE = 13290,49$; $RMSE = 22213,76$; $R^2 = 0,9385$). Наступним етапом дослідження стала інтерпретація отриманих прогнозів, що дозволяє визначити, які характеристики об'єктів нерухомості найбільше впливають на оцінку їх вартості та яким чином модель формує прогноз.

Для аналізу важливості ознак спочатку було використано вбудований механізм оцінювання Feature Importance моделі CatBoost. Результати наведено на рисунку 4.10.

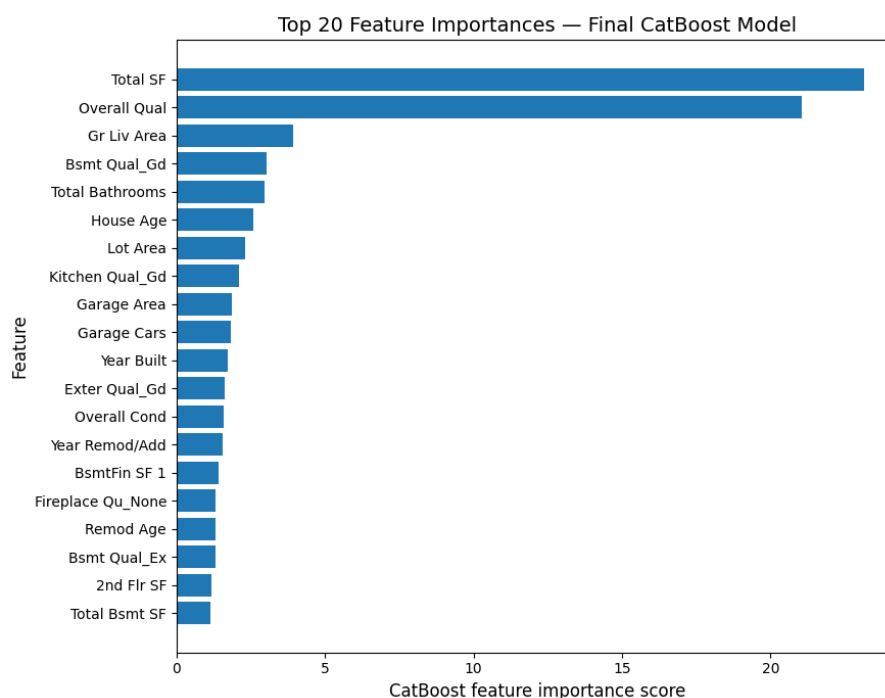


Рисунок 4.10 – Найважливіші ознаки фінальної моделі CatBoost

Як видно з рисунка 4.10, найбільший вплив на прогнозовану вартість об'єктів має ознака Total SF, яка характеризує сумарну площу будинку та була сформована на етапі feature engineering. Друге місце за важливістю посідає Overall Qual, що відображає загальну якість будівництва та оздоблення. Серед найбільш значущих також присутні Gr Liv Area, Total Bathrooms, House Age, Lot Area, Garage Area, Garage Cars, Year Built та інші характеристики, які безпосередньо описують площу, технічний стан і функціональні властивості житлового будинку.

Отримані результати підтверджують доцільність сформованих додаткових ознак. Зокрема, агрегована ознака Total SF виявилася найінформативнішою серед усіх характеристик, що свідчить про ефективність застосованого етапу feature engineering.

Для порівняння результатів моделі із статистичними залежностями між ознаками та цільовою змінною було проведено кореляційний аналіз числових характеристик датасету. Найбільш корельовані ознаки наведено на рисунку 4.11.

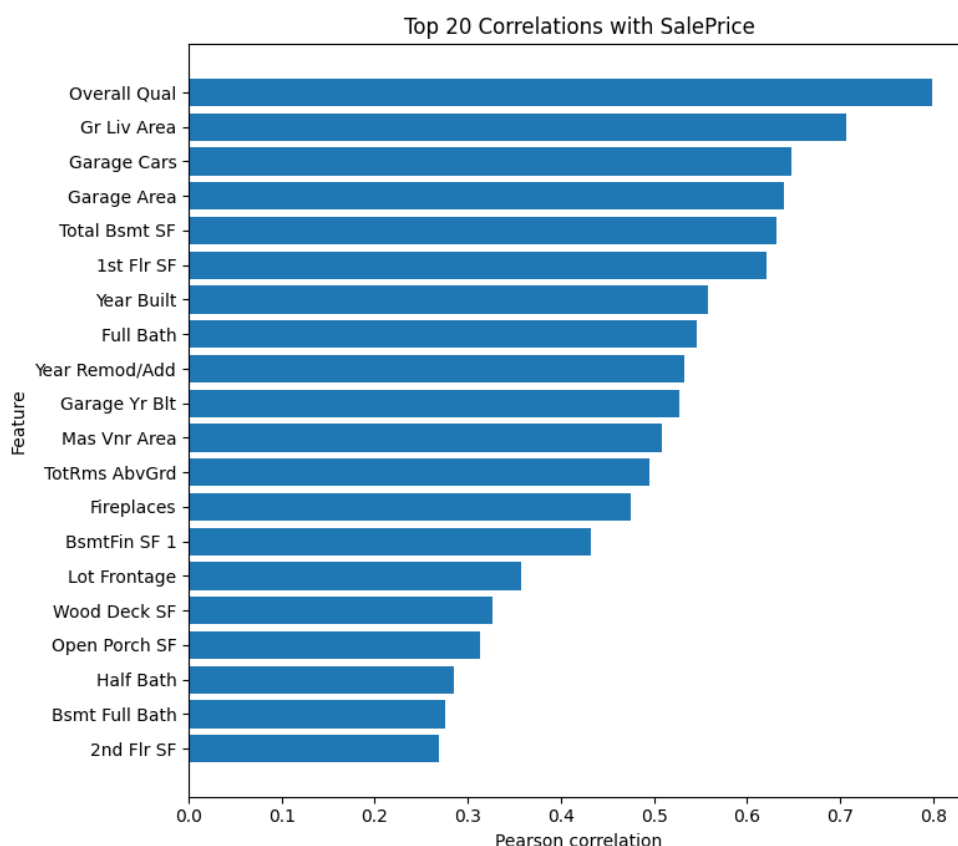


Рисунок 4.11 – Кореляція числових ознак із вартістю об'єктів нерухомості

Кореляційний аналіз показав, що найбільший лінійний зв'язок із вартістю об'єктів мають ознаки Overall Qual, Gr Liv Area, Garage Cars, Garage Area, Total Bsmt SF та 1st Flr SF. Водночас результати частково відрізняються від Feature Importance моделі CatBoost. Наприклад, ознака Total SF, яка є найважливішою для моделі, відсутня серед початкових характеристик із найбільшою кореляцією, оскільки була сформована на етапі побудови нових ознак. Це свідчить про те, що модель використовує не лише окремі характеристики, а й більш інформативні агреговані ознаки та враховує нелінійні взаємозв'язки між ними.

Для детальнішого пояснення роботи моделі було застосовано метод SHAP (SHapley Additive exPlanations), який дозволяє оцінити внесок кожної ознаки у формування прогнозу. На рисунку 4.12 наведено SHAP Summary Plot.

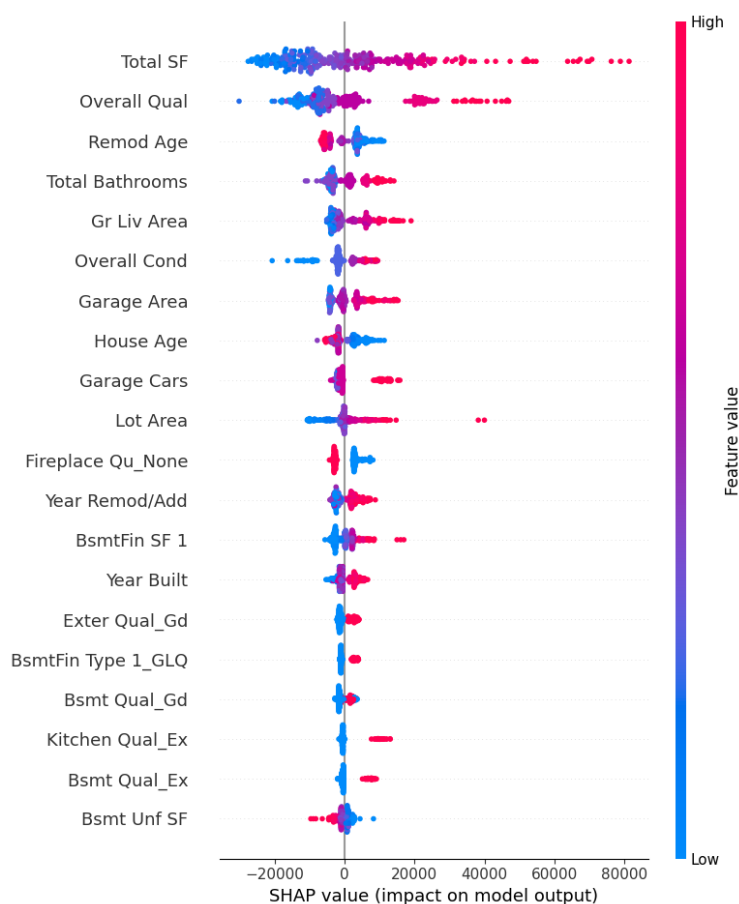


Рисунок 4.12 – SHAP Summary Plot для фінальної моделі CatBoost

SHAP Summary Plot відображає одночасно важливість ознак, напрямок їх впливу та розподіл внеску для всіх досліджуваних об'єктів. Кожна точка відповідає окремому будинку, а її положення по горизонталі характеризує величину внеску ознаки у прогнозовану вартість. Колір точки відображає значення відповідної характеристики: синій колір відповідає низьким значенням ознаки, а червоний – високим.

Найбільший внесок у прогноз мають Total SF та Overall Qual. Для цих ознак високі значення переважно зміщують прогноз у бік збільшення вартості нерухомості, тоді як низькі значення, навпаки, зменшують прогнозовану ціну. Аналогічна закономірність спостерігається для Gr Liv Area, Garage Area, Garage Cars та Total Bathrooms. Для ознаки House Age характер залежності є протилежним: збільшення віку будинку переважно знижує прогнозовану ринкову вартість.

Середню абсолютну величину впливу кожної ознаки наведено на рисунку 4.13.

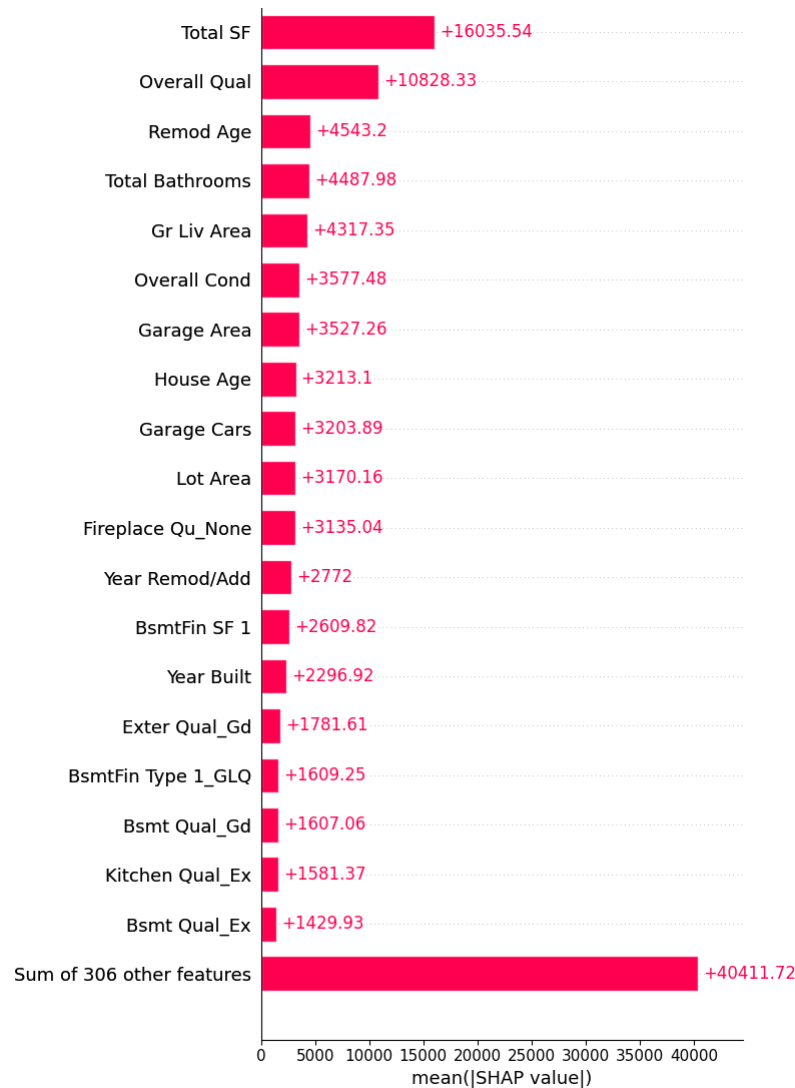


Рисунок 4.13 – SHAP Bar Plot для фінальної моделі CatBoost

SHAP Bar Plot підтверджує результати попереднього аналізу та ранжує ознаки за середнім абсолютним внеском у прогноз моделі. Найбільший вплив мають Total SF, Overall Qual, Remod Age, Total Bathrooms та Gr Liv Area, що узгоджується як із результатами Feature Importance, так і з SHAP Summary Plot.

Для пояснення прогнозу окремого об'єкта нерухомості побудовано SHAP Waterfall Plot, наведений на рисунку 4.14.

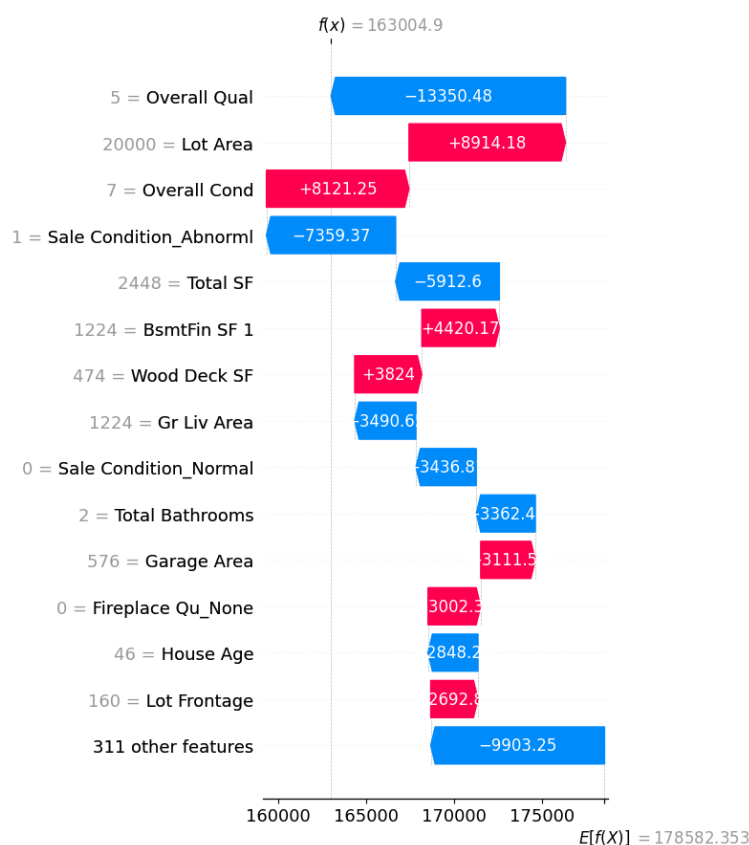


Рисунок 4.14 – SHAP Waterfall Plot для окремого прогнозу моделі CatBoost

Waterfall Plot демонструє процес формування прогнозу для конкретного будинку. Побудова прогнозу починається із середнього значення вартості нерухомості в навчальній вибірці, після чого кожна ознака послідовно збільшує або зменшує це значення. На досліджуваному прикладі найбільший позитивний внесок забезпечують Lot Area, Overall Cond, BsmtFin SF 1 та Garage Area, тоді як найбільше знижують прогноз Overall Qual, Sale Condition, Total SF та деякі інші характеристики. Сума всіх індивідуальних внесків формує кінцеве прогнозоване значення вартості об'єкта.

Таким чином, результати інтерпретації підтвердили, що фінальна модель CatBoost приймає рішення на основі характеристик, які безпосередньо визначають споживчу цінність нерухомості: площі будинку, загальної якості будівництва, кількості санвузлів, віку будівлі, параметрів

земельної ділянки та інших конструктивних особливостей. Використання SHAP дозволило не лише визначити найважливіші фактори, але й пояснити їхній внесок як на рівні всієї моделі, так і для окремих прогнозів, що підвищує прозорість та інтерпретованість розробленої інтелектуальної системи.

4.5 Аналіз точності прогнозування

Після вибору фінальної моделі CatBoost було проведено детальний аналіз похибок прогнозування. Його метою є оцінювання характеру помилок моделі, визначення особливостей прогнозування для різних категорій об'єктів нерухомості та виявлення випадків, у яких модель демонструє найбільші відхилення від фактичної вартості.

Спочатку проаналізовано загальну якість прогнозування фінальної моделі. Для цього побудовано графік відповідності прогнозованих і фактичних значень, графік залишків та гістограму розподілу залишків, які наведено на рисунках 4.15–4.17.

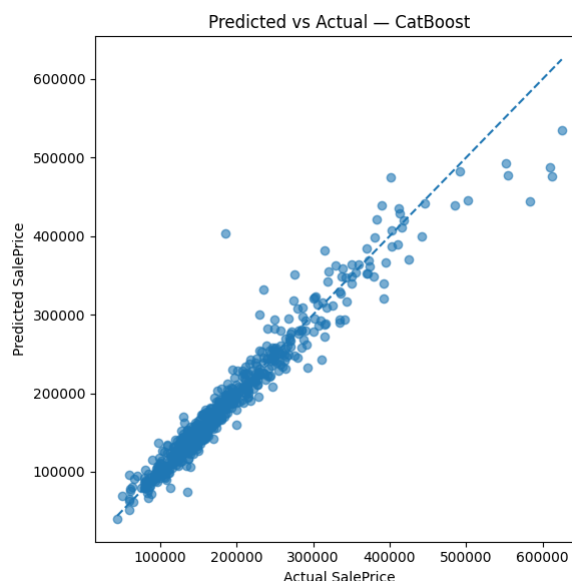


Рисунок 4.15 – Порівняння фактичної та прогнозованої вартості об'єктів фінальною моделлю CatBoost

Як видно з рисунка 4.16, більшість точок розташована поблизу діагональної лінії, що відповідає ідеальному прогнозу. Це свідчить про високу узгодженість між фактичними та прогнозованими значеннями вартості нерухомості. Найбільші відхилення спостерігаються для об'єктів із дуже високою ціною, тоді як для основної частини вибірки модель забезпечує достатньо точне прогнозування.

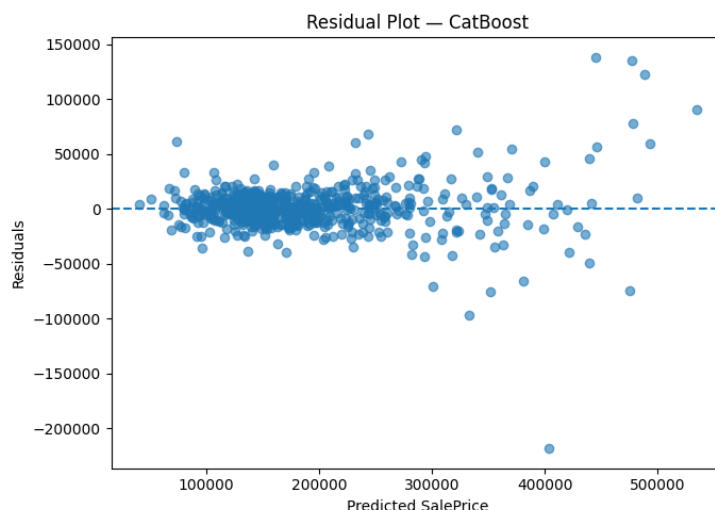


Рисунок 4.16 – Графік залишків фінальної моделі CatBoost

Графік залишків, наведений на рисунку 4.16, демонструє, що більшість похибок зосереджена поблизу нульового значення без вираженого систематичного зміщення. Водночас зі збільшенням прогнозованої вартості нерухомості спостерігається поступове зростання розкиду залишків. Це свідчить про те, що прогнозування дорогих об'єктів є складнішим порівняно з масовим сегментом ринку, що є типовою особливістю задач прогнозування вартості нерухомості.

Гістограма залишків (рисунок 4.17) підтверджує, що переважна більшість похибок концентрується поблизу нульового значення та має близький до симетричного розподіл. Наявні окремі великі позитивні та негативні відхилення відповідають невеликій кількості нетипових об'єктів із максимальною похибкою прогнозування.

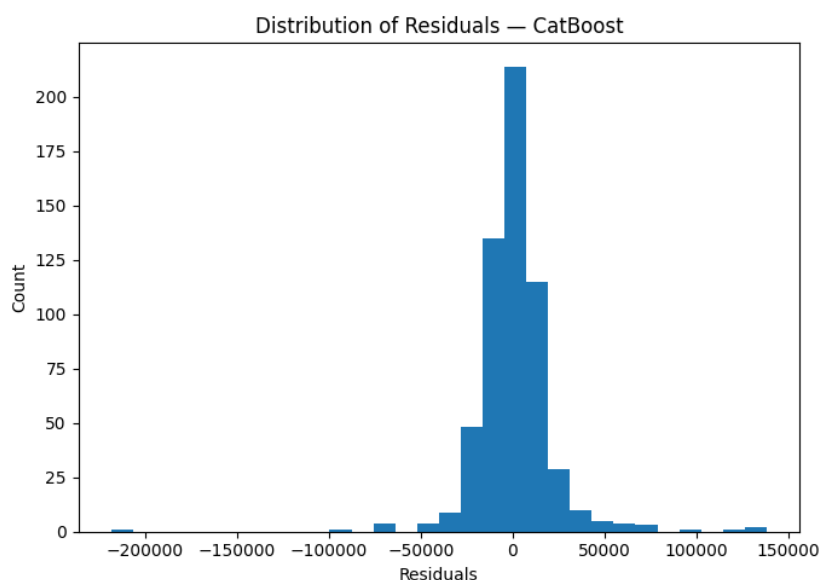


Рисунок 4.17 – Розподіл залишків фінальної моделі CatBoost

Після загальної оцінки роботи моделі виконано аналіз похибок залежно від вартості нерухомості. Для цього тестову вибірку було поділено на чотири цінові сегменти за допомогою кuartильного розбиття фактичних значень цільової ознаки. Для кожного сегмента обчислено середню абсолютну похибку (MAE), середньоквадратичну похибку (RMSE), а також середні фактичні та прогнозовані значення вартості. Отримані результати наведено на рисунку 4.18.

	Price Segment	Count	Min_Price	Max_Price	MAE	RMSE	Mean_Actual	Mean_Predicted
0	Low price	148	44000	130000	8772.474749	11480.660096	106105.351351	109135.954143
1	Medium-low price	147	130500	165000	9269.910201	12370.396679	146402.210884	145547.129668
2	Medium-high price	144	165500	228000	11463.320617	22107.495688	194136.826389	196164.743927
3	High price	147	228500	625000	23649.686166	34678.441145	312621.367347	305307.662659

Рисунок 4.18 – Показники похибки прогнозування за ціновими сегментами

Для кращої візуалізації результатів побудовано діаграму середньої абсолютної похибки для кожного цінового сегмента (рисунок 4.19).

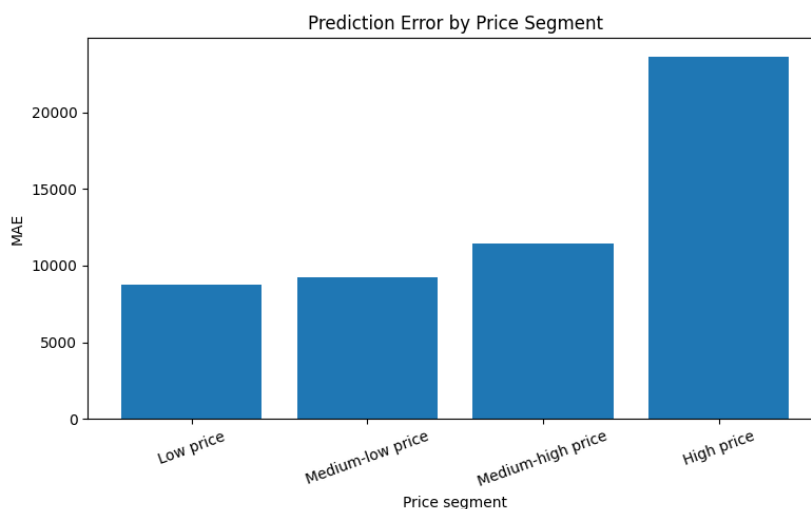


Рисунок 4.19 – Середня абсолютна похибка прогнозування за ціновими сегментами

Отримані результати показують, що найменшу похибку модель демонструє для об'єктів нижнього цінового сегмента, де середнє значення MAE становить близько 8,8 тис. грошових одиниць. Для сегмента Medium-low price цей показник становить приблизно 9,3 тис., для Medium-high price – 11,5 тис., тоді як для найдорожчих будинків середня абсолютна похибка зростає до 23,6 тис. грошових одиниць.

Подібна закономірність пояснюється тим, що дорогі об'єкти нерухомості суттєво відрізняються між собою за площею, плануванням, якістю будівництва, рівнем оздоблення та іншими характеристиками. Крім того, кількість таких об'єктів у навчальній вибірці є значно меншою, ніж об'єктів середнього цінового сегмента, що ускладнює процес навчання моделі. Водночас навіть для дорогих будинків відносна похибка прогнозування залишається невеликою порівняно з їх загальною вартістю.

Для більш детального дослідження виконано аналіз двадцяти об'єктів із найбільшою абсолютною похибкою прогнозування. Результати наведено на рисунку 4.20.

Додатково досліджено розподіл цих об'єктів між ціновими сегментами. Отримані результати наведено на рисунку 4.21.

	Actual SalePrice	Predicted SalePrice	Residual	Absolute Error	Neighborhood	Overall Qual	Gr Liv Area	Total SF	Garage Cars	Total Bathrooms	Year Built	House Age
2181	184750	403310.484181	-218560.484181	218560.484181	Edwards	10	4676	7814.0	3.0	4.5	2007	0
433	582933	444768.988095	138164.011905	138164.011905	NridgHt	9	2822	4556.0	3.0	3.5	2008	1
44	611657	476631.611916	135025.388084	135025.388084	NridgHt	9	2364	4694.0	3.0	3.5	2009	1
432	610000	487848.467817	122151.532183	122151.532183	NridgHt	10	2674	5304.0	3.0	4.5	2007	2
2246	235000	332247.817418	-97247.817418	97247.817418	Crawfor	8	1625	3198.0	2.0	3.5	2006	1
2445	625000	534230.261067	90769.738933	90769.738933	NoRidge	10	3627	5557.0	3.0	4.5	1995	11
423	555000	477596.067941	77403.932059	77403.932059	NridgHt	10	2402	5496.0	3.0	3.0	2008	1
2103	276000	351220.769150	-75220.769150	75220.769150	CollgCr	8	2574	4056.0	3.0	2.5	2007	0
2335	401179	475343.894653	-74164.894653	74164.894653	StoneBr	9	2795	4185.0	3.0	3.5	2005	1
1559	392500	321031.465162	71468.534838	71468.534838	Crawfor	8	1652	3252.0	2.0	3.5	2008	0
944	230000	300459.241207	-70459.241207	70459.241207	Crawfor	8	1648	3244.0	2.0	3.5	2007	2
1538	311500	243174.087258	68325.912742	68325.912742	Crawfor	7	1954	2752.0	2.0	2.0	1939	69
2378	315000	381132.214288	-66132.214288	66132.214288	NridgHt	9	1980	3960.0	3.0	3.0	2004	2
2620	135000	73809.751442	61190.248558	61190.248558	NAmes	4	1192	1912.0	0.0	1.5	1949	57
2383	293000	232293.015029	60706.984971	60706.984971	NridgHt	7	1740	3461.0	3.0	2.0	2006	0
456	552000	492493.933204	59506.066796	59506.066796	NridgHt	10	2492	4984.0	3.0	3.5	2004	5
366	501837	445723.728592	56113.271408	56113.271408	StoneBr	9	2234	4450.0	3.0	3.0	2008	1
2097	424870	370437.926387	54432.073613	54432.073613	CollgCr	8	2828	4294.0	3.0	4.0	2006	1
2275	392000	340009.649452	51990.350548	51990.350548	Timber	10	1824	3648.0	3.0	2.0	2006	1
1058	390000	439236.341042	-49236.341042	49236.341042	NridgHt	9	2748	4598.0	3.0	3.5	2006	2

Рисунок 4.20 – Об'єкти з найбільшою абсолютною похибкою прогнозування

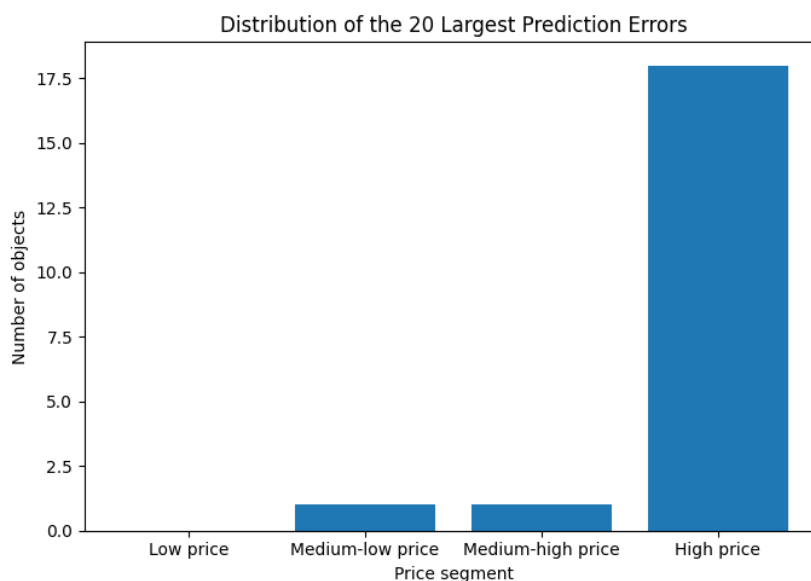


Рисунок 4.21 – Розподіл двадцяти найбільших похибок прогнозування за ціновими сегментами

Як видно з рисунка 4.21, переважна більшість найбільших похибок припадає саме на сегмент High price. Із двадцяти об'єктів вісімнадцять належать до найдорожчої категорії, тоді як лише по одному випадку

зафіксовано для сегментів Medium-low price та Medium-high price. Для найдешевших об'єктів значних похибок прогнозування не виявлено.

Аналіз рисунку 4.20 показує, що найбільші відхилення виникають переважно для будинків із дуже великою площею, високими значеннями показника Overall Qual, значною кількістю санвузлів, великими гаражами та іншими нетиповими характеристиками. Частина таких об'єктів також належить до престижних районів забудови, де ринкова вартість визначається не лише характеристиками, представленими у наборі Ames Housing Dataset, а й додатковими чинниками, які відсутні у вихідних даних. Це обмежує можливість моделі повністю врахувати всі особливості формування ціни.

Отже, проведений аналіз підтвердив, що фінальна модель CatBoost забезпечує стабільно високу точність прогнозування для більшості об'єктів нерухомості та демонструє найкращі результати у масовому сегменті ринку. Основні похибки зосереджені серед найдорожчих будинків, що пояснюється високою різноманітністю їх характеристик та обмеженим представленням таких об'єктів у навчальній вибірці.

ВИСНОВКИ

У кваліфікаційній роботі вирішено актуальне завдання розробки інтелектуальної системи прогнозування вартості об'єктів нерухомості на основі методів машинного навчання. У межах роботи реалізовано експериментальний ML pipeline, який охоплює всі основні етапи побудови моделі прогнозування: формування експериментального набору даних, попередню обробку, формування додаткових ознак, навчання моделей, оцінювання їх якості, налаштування гіперпараметрів та інтерпретацію отриманих результатів.

У першому розділі досліджено предметну галузь прогнозування вартості нерухомості, проаналізовано фактори, що визначають ринкову вартість житлових об'єктів, та виконано огляд сучасних інформаційних систем і підходів до автоматизованої оцінки нерухомості. Проведений аналіз показав, що використання методів машинного навчання дозволяє враховувати складні взаємозв'язки між характеристиками об'єктів та підвищувати точність прогнозування порівняно з традиційними підходами.

У другому розділі проаналізовано сучасні методи машинного навчання, що застосовуються для задач регресії, обґрунтовано вибір моделей Linear Regression, Random Forest, Extra Trees, XGBoost та CatBoost для проведення експериментального дослідження. Розглянуто принципи побудови ML pipeline, метрики оцінювання якості моделей (MAE, RMSE та R^2), а також сучасні методи інтерпретації моделей машинного навчання на основі SHAP.

У третьому розділі розроблено архітектуру інтелектуальної системи прогнозування та реалізовано програмний ML pipeline експериментального дослідження. Використано набір даних Ames Housing Dataset, проведено аналіз його структури та якості, сформовано навчальну і тестову вибірки. Реалізовано базову та покращену схеми попередньої обробки даних, які включають імпутацію пропущених значень, кодування категоріальних

ознак та використання єдиного конвеєра обробки даних. Запропоновано набір додаткових ознак, що характеризують площу будинку, його вік, кількість санвузлів, площу прибудов та інші конструктивні характеристики. Також реалізовано універсальні програмні компоненти для навчання моделей, автоматизованого оцінювання їх якості та збереження результатів експериментів.

У четвертому розділі проведено комплексне експериментальне дослідження реалізованого ML pipeline. Встановлено, що вдосконалена схема попередньої обробки даних забезпечує покращення якості прогнозування порівняно з базовим варіантом. Дослідження впливу feature engineering показало, що сформовані агреговані ознаки позитивно впливають на ансамблеві моделі, тоді як для лінійної регресії їх використання не забезпечує підвищення точності. Порівняльний аналіз моделей продемонстрував перевагу ансамблевих методів і моделей градієнтного бустингу над класичною лінійною регресією.

Для моделей XGBoost та CatBoost виконано налаштування гіперпараметрів із використанням RandomizedSearchCV, що дозволило додатково покращити якість прогнозування. Найкращі результати отримано для моделі CatBoost після оптимізації гіперпараметрів: $MAE = 13290,49$, $RMSE = 22213,76$ та $R^2 = 0,9385$. За сукупністю отриманих результатів саме ця модель була визначена як найбільш ефективна для досліджуваної задачі.

Виконано інтерпретацію прогнозів фінальної моделі за допомогою Feature Importance та методології SHAP. Встановлено, що найбільший вплив на прогнозовану вартість мають сумарна площа будинку, загальна якість будівництва, житлова площа, кількість санвузлів, площа земельної ділянки, параметри гаража та вік будинку. Проведений аналіз підтвердив доцільність використання сформованих додаткових ознак і дозволив пояснити внесок окремих характеристик як на рівні всієї моделі, так і для конкретних прогнозів.

Додатковий аналіз похибок прогнозування показав, що найвища точність досягається для об'єктів нижнього та середнього цінових сегментів, тоді як найбільші відхилення спостерігаються для дорогих будинків із нетиповими характеристиками. Встановлено, що основні помилки пов'язані з обмеженим представленням таких об'єктів у навчальній вибірці та високою варіативністю факторів, які визначають їх ринкову вартість.

Практичне значення роботи полягає у реалізації програмного ML pipeline, який може бути використаний як основа для подальшої розробки інтелектуальних систем прогнозування вартості нерухомості. Запропонований підхід є модульним, що дозволяє розширювати набір ознак, інтегрувати нові моделі машинного навчання та адаптувати його до інших наборів даних і практичних задач оцінювання нерухомості.

Отримані результати підтверджують ефективність комплексного підходу, що поєднує попередню обробку даних, формування нових ознак, використання ансамблевих моделей, оптимізацію гіперпараметрів та сучасні методи інтерпретації машинного навчання. Реалізований ML pipeline може слугувати основою для подальших досліджень і розвитку інтелектуальних систем підтримки прийняття рішень у сфері оцінювання нерухомості.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Піта Ю. А. Ринок нерухомості як складник української економіки: особливості стимулюючого впливу. *Трансформаційна економіка*. 2025. № 3 (12). С. 69–78. URL: <https://doi.org/10.32782/2786-8141/2025-12-11> (дата звернення: 08.06.2026).
2. Як визначити ринкову вартість нерухомості. *DOM.RIA.com*. URL: <https://dom.ria.com/uk/articles/kak-opredelit-rynochnuyu-stoimost-nedvizhimosti-254388.html> (дата звернення: 28.06.2026).
3. Статистичний аналіз ролі основних детермінант у формуванні вартості активів житлової нерухомості в Україні. *AFO*. URL: <https://afo.com.ua/uk/news/2/1512> (дата звернення: 08.06.2026).
4. Sotnikov O. The impact of macroeconomic indicators on mortgage lending for residential real estate. *Scientific bulletin of Odessa national economic university*. 2024. Vol. 9, no. 322. P. 54–62. URL: <https://doi.org/10.32680/2409-9260-2024-9-322-54-62> (date of access: 09.06.2026).
5. Veres O., Shymoniak A. Forecasting the value of real estate using machine learning tools. *Visnik Natsional'nogo universitetu "L'vivs'ka politehnika". Seriâ Informacijni sistemi ta mereži*. 2024. Vol. 15. P. 140–158. URL: <https://doi.org/10.23939/sisn2024.15.140> (date of access: 09.06.2026).
6. Selim H. Determinants of house prices in Turkey: hedonic regression versus artificial neural network. *Expert systems with applications*. 2009. Vol. 36, no. 2. P. 2843–2852. URL: <https://doi.org/10.1016/j.eswa.2008.01.044> (date of access: 09.06.2026).
7. Park B., Bae J. K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications*. 2015. Vol. 42, no. 6. P. 2928–2934. URL: <https://doi.org/10.1016/j.eswa.2014.11.040> (date of access: 09.06.2026).
8. Li H. House price prediction and analysis based on random forest and xgboost models. *Highlights in business, economics and management*. 2023.

Vol. 21. P. 934–938. URL: <https://doi.org/10.54097/hbem.v21i.14837> (date of access: 09.06.2026).

9. Zou C. The house price prediction using machine learning algorithm: the case of Jinan, China. *Highlights in science, engineering and technology*. 2023. Vol. 39. P. 327–333. URL: <https://doi.org/10.54097/hset.v39i.6549> (date of access: 10.06.2026).

10. Housing market prediction problem using different machine learning algorithms: a case study / S. B. Jha et al. *arXiv*. 2020. URL: <https://doi.org/10.48550/arXiv.2006.10092> (date of access: 10.06.2026).

11. Bourassa S., Cantoni E., Hoesli M. Predicting house prices with spatial dependence: a comparison of alternative methods. *Journal of real estate research*. 2010. Vol. 32, no. 2. P. 139–160. URL: <https://doi.org/10.1080/10835547.2010.12091276> (date of access: 10.06.2026).

12. Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms / A. Soltani et al. *Cities*. 2022. Vol. 131. P. 103941. URL: <https://doi.org/10.1016/j.cities.2022.103941> (date of access: 10.06.2026).

13. Kim W., Hong J. Stacked ensemble model for the automatic valuation of residential properties in south korea: a case study on jeju island. *Land*. 2024. Vol. 13, no. 9. P. 1436. URL: <https://doi.org/10.3390/land13091436> (date of access: 10.06.2026).

14. Acharya D. B., Divya B., Kuppan K. Explainable and fair AI: balancing performance in financial and real estate machine learning models. *IEEE Access*. 2024. P. 1. URL: <https://doi.org/10.1109/access.2024.3484409> (date of access: 10.06.2026).

15. Lee F. Classification vs regression. *IBM*. URL: <https://www.ibm.com/think/topics/classification-vs-regression> (date of access: 11.06.2026).

16. Linear regression in machine learning. *GeeksforGeeks*.
URL: <https://www.geeksforgeeks.org/machine-learning/ml-linear-regression/>
(date of access: 11.06.2026).
17. Understanding decision tree regressor: an in-depth intuition. *Medium*.
URL: <https://farshadabdulazeez.medium.com/understanding-decision-tree-regressor-an-in-depth-intuition-a1d3af182efd> (date of access: 11.06.2026).
18. Murel J., Kavlakoglu E. What is ensemble learning?. *IBM*.
URL: <https://www.ibm.com/think/topics/ensemble-learning> (date of access: 11.06.2026).
19. Akturk M. What is the “boosting” ensemble method?. *Medium*.
URL: <https://mathchi.medium.com/what-is-the-boosting-ensemble-method-76610a5cb39f> (date of access: 11.06.2026).
20. Kavlakoglu E. What is random forest?. *IBM*.
URL: <https://www.ibm.com/think/topics/random-forest> (date of access: 12.06.2026).
21. Clark B., Lee F. What is gradient boosting? | IBM. *IBM*.
URL: <https://www.ibm.com/think/topics/gradient-boosting> (date of access: 12.06.2026).
22. Pandelu A. P. XGBoost and catboost – optimized gradient boosting models. *Medium*. URL: <https://medium.com/@bhatadithya54764118/day-19-xgboost-and-catboost-optimized-gradient-boosting-models-c58b57bdc109> (date of access: 12.06.2026).
23. Jones D. Evaluation metrics for regression models: a practical guide. *Medium*. URL: <https://medium.com/@duygujones/evaluation-metrics-for-regression-models-a-practical-guide-3e5d27ef628c> (date of access: 12.06.2026).
24. An introduction to explainable AI with Shapley values . *Welcome to the SHAP documentation – SHAP latest documentation*.
URL: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html
1 (date of access: 12.06.2026).

25. Ames housing dataset. *Kaggle: The World's AI Proving Ground*.

URL: <https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset> (date of access: 13.06.2026).