КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій
Кафедра технологій управління

Спеціальність 122 – Комп'ютерні науки,
освітня програма «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

" Інформаційний аналіз та прогнозування даних онлайн-сервісів ринку нерухомості"

Студента 2-го курсу групи ІАВ-21

Склярова Андрія Олександровича
(прізвище, ім'я, по батькові)

(підпис студента)

Науковий керівник:

Кандидат технічних наук, асистент
(науковий ступінь, вчене звання)

Хлевний Андрій Олександрович
(прізвище, ім'я, по батькові)

(дата) (підпис)

Попередній захист:

(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри технологій управління

(підпис) (прізвище, ініціали) (дата)

Київ – 2023

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій

Кафедра технологій управління
Освітньо-кваліфікаційний рівень Магістр
Спеціальність 122 - Комп'ютерні науки
Освітня програма Інформаційна аналітика та впливи

ЗАТВЕРДЖУЮ
Завідувач кафедри
професор Морозов В.В.

_____
«____» _____20___року

З А В Д А Н Н Я
НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студент Скляров Андрій Олександрович
Група___ІАВ-21_____

1. Тема кваліфікаційної роботи
Інформаційний аналіз та прогнозування даних онлайн-сервісів ринку нерухомості_____

Затверджена наказом по від «___»_____20__ р. № ___.
2. Строк подання студентом готової роботи – "___"_____ 20___ р.

3. Цільова установка та вихідні дані до роботи: дослідження теоретичної бази сфери ринку нерухомості та застосування інформаційної аналітики у даній сфері; аналіз методів та методик аналізу та моделювання процесів, включно з програмними засобами реалізації; побудова прогностичної моделі та її оцінка; проведення інформаційної розвідки щодо потенційної імплементації отриманої моделі на підприємствах.

4. Зміст роботи: аналіз ринку нерухомості, інформаційна аналітика у сфері продажу нерухомості, методи інформаційного аналізу та прогнозування даних, математичні методи для моделювання, вибір даних для аналізу ринку нерухомості, обробка даних, побудова моделі методом лінійної регресії, побудова моделі методом регуляризації, побудова моделі методом Random forest, побудова моделі методом XGBoost, порівняння ефективності моделей, порівняння результатів роботи моделей, застосування розробленої моделі на підприємствах комерційного спрямування, потенціал розвитку проєкту.

5. Перелік графічного матеріалу (слайдів) Динаміка обсягів нової пропозиції на ринку житлової нерухомості України, 2022-й до 2021 року; Динаміка кількості укладених угод купівлі-продажу житла на вторинному ринку України, 2022-й до 2021 року; Динаміка середніх цін (в доларах США) на оренду та купівлю квартир по областях, грудень 2022-го до грудня 2021 року; Кореляційна матриця; Структура даних; Графіки відношення залежних і незалежних змінних; Результати роботи моделей

6. Календарний план виконання роботи:

№ з/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1	Вибір теми дипломної роботи	3	08.12.2022	08.12.2022
2	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	08.12.2022	08.12.2022
3	Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи	10	08.01.2023	08.01.2023
4	Складання розгорнутого плану кваліфікаційної роботи	5	18.01.2023	18.01.2023
5	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін.	5	20.01.2023	20.01.2023
6	Підготовка розділу 1 “Теоретичні основи цінового прогнозування на житлову нерухомість”	10	13.02.2023	13.02.2023
7	Підготовка розділу 2 “Аналіз методів та методологій інформаційного аналізу та моделювання”	15	06.03.2023	06.03.2023
8	Підготовка розділу 3 “Практичне застосування описаних методів для побудови прогнозу”	20	03.04.2023	03.04.2023
9	Підготовка розділу 4 “Практична цінність та можливості використання”	13	17.04.2023	17.04.2023
10	Оформлення кваліфікаційної роботи. Підготовка висновків і пропозицій	12	01.05.2023	01.05.2023
11	Передача кваліфікаційної роботи науковому керівникові	2	02.05.2023	02.05.2023
12	Передача кваліфікаційної роботи рецензенту для рецензування	1	10.05.2023	10.05.2023
13	Попередній захист кваліфікаційної роботи	2	17.05.2023	17.05.2023

Дата видачі завдання «____» _____ 20__ р.

Керівник роботи _к.т.н., асистент Хлевний Андрій Олександрович_____
(посада, прізвище, ім’я, по батькові)

_____
(підпис)

Завдання прийняв до виконання студент групи _ІАВ-21____
___Скляров Андрій Олександрович_____
(прізвище, ім’я, по батькові)

_____
(підпис)

ЗМІСТ

АНОТАЦІЯ

Київський національний університет імені Тараса Шевченка
Факультет інформаційних технологій
Кафедра технологій управління
Спеціальність 122 – Комп’ютерні науки,
освітня програма «Інформаційна аналітика та впливи»

Дипломна робота магістранта Склярова Андрія Олександровича.

Тема роботи – «Інформаційний аналіз та прогнозування даних онлайн-сервісів ринку нерухомості».

Мета дипломної роботи магістра – проаналізувати ринок нерухомості України та розробити модель прогнозування цін на житло за допомогою математичних методів машинного навчання.

Об’єкт дослідження. ціна на житлову нерухомість.

Предмет дослідження. Математичні методи моделювання і прогнозування ціни на житлову нерухомість, їх аналіз та оцінка.

Наукова новизна полягає у визначенні важливих аспектів функціонування ринку житлової нерухомості, застосуванні алгоритмів машинного навчання для аналізу великої кількості даних про ринок нерухомості, розробці ефективної моделі машинного навчання, що допомогає виявляти складні зв'язки та тренди, що впливають на ціни.

Для досліджень теоретичних аспектів житлового ринку були використані методи дедукції та індукції, аналізу та синтезу. Для аналізу та прогнозування ціни було застосовано системний та комплексний підхід, статистичні методи збору та обробки інформації. Для прогнозування було використано чотири методи – лінійна регресія, регуляризація, випадковий ліс та XGBoost. Визначено найбільш точну з них та обґрунтування результатів з економічної точки зору.

Дипломна робота складається зі вступу, основної частини, що включає 4 розділи, висновку, списку використаних джерел та додатків. Всього налічує 148 сторінку, перелік з 78 джерел на 10 сторінках та 6 додатків на 23 сторінках.

Ключові слова: машинне навчання, методи, інформаційна аналітика даних, ринок нерухомості, моделювання, аналіз ринку нерухомості, прогнозування цін.

ВСТУП

Актуальність роботи. Ринок нерухомості грає важливу роль у економіці, встановлюючи зв'язок між будівельною галуззю та споживачами житла, визначаючи ціни на нерухомість та найбільш ефективні форми житла, перерозподіляючи житло серед громадян та встановлюючи взаємозв'язок між попитом та пропозицією на житло. Більше того, групи, зацікавлені у розумінні сучасного стану та прогнозуванні тенденцій, не обмежуються лише орендарями та орендодавцями, покупцями та продавцями, забудовниками або брокерами, а також включають приватних інвесторів, банки, фірми, страхові компанії, державу.

Оскільки ринок нерухомості має велике значення для економіки країни, структурний аналіз та активний розвиток цього ринку є важливими напрямками досліджень. Прогнозування цін на нерухомість є актуальною та важливою задачею, яку можна вирішити з використанням методів машинного навчання. Дані методи стали первинним джерелом вдосконалених можливостей для консолідації, аналізу, прогнозування, звітності та візуалізації цін на нерухомість.

Метою роботи є проаналізувати ринок нерухомості України та розробити модель прогнозування цін на житло за допомогою математичних методів машинного навчання. Дана мета реалізується через виконання наступних завдань:

1. Опрацювання теоретичних засади ринку житлової нерухомості;
2. Визначити фактори впливу на ціну нерухомості;
3. Проаналізувати сучасний стан та тенденції ринку нерухомості України;
4. Визначити основні математичні алгоритми, які можна застосувати для прогнозування ціни житлової нерухомості;
5. Побудувати моделі прогнозування ціни та обрати найбільш оптимальну з економічної точки зору.

Об’єктом дослідження є ціна на житлову нерухомість.

Предметом дослідження є математичні методи моделювання і прогнозування ціни на житлову нерухомість, їх аналіз та оцінка.

Методологія та методи дослідження. Методологічною основою дослідження послужили теоретичні напрацювання українських та зарубіжних вчених із питань, пов’язаних із розвитком ринку житлової нерухомості. Для досліджень теоретичних аспектів житлового ринку були використані методи дедукції та індукції, аналізу та синтезу. Для аналізу та прогнозування ціни було застосовано системний та комплексний підхід, статистичні методи збору та обробки інформації. Для прогнозування було використано чотири методи – лінійна регресія, регуляризація, випадковий ліс та XGBoost.

Теоретична, методична та практична значущість роботи полягає у визначенні важливих аспектів функціонування ринку житлової нерухомості. З отриманих результатів, фахівці зможуть краще планувати фінанси, інвестиції, проводити економічний аналіз ринку. Практична значущість відображається у можливості використання моделі машинного навчання з метою впровадження її на підприємстві як допоміжного інструменту аналізу даних.

Наукова новизна полягає у застосуванні алгоритмів машинного навчання для аналізу великої кількості даних про ринок нерухомості, розробці ефективної моделі машинного навчання, що допомогає виявляти складні зв'язки та тренди, що впливають на ціни. Зокрема, було побудовано 4 моделі для прогнозування ціни на будинки, визначено найбільш точну з них та обґрунтування результатів з економічної точки зору.

Інформаційною базою дослідження послужили праці вітчизняних та зарубіжних науковців і практиків, наукові періодичні видання, статистичні дані спеціалізованих інтернет-журналів.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ ЦІНОВОГО ПРОГНОЗУВАННЯ НА ЖИТЛОВУ НЕРУХОМІСТЬ

1.1 Сутність та специфіка ринку житлової нерухомості

Нерухомість – це форма власності, що охоплює земельну ділянку та забудову, включаючи будівлі, споруди та інші непересувні матеріальні об'єкти [1].

Можна сказати, що нерухомість охоплює фізичну площу землі, її поверхню, що знаходиться над і під нею, постійно прив'язану до неї, а також всі права власності, такі як право володіння, продажу, оренди та користування землею. Існує п'ять основних типів нерухомості:

1. Житлова нерухомість - це будь-яке майно, яке використовується для проживання. Це можуть бути будинки, квартири, котеджі, кімнати в багатоквартирних будинках та інші.
2. Комерційна нерухомість - це будь-яке майно, яке використовується виключно для комерційних цілей, наприклад, житлові комплекси, заправні станції, продуктові магазини, лікарні, готелі, офіси, паркінги, ресторани, торгові центри, магазини та театри.
3. Промислова нерухомість - це будь-яке майно, яке використовується для виробництва, розподілу, зберігання, досліджень та розробок. Це можуть бути заводи, електростанції та склади.
4. Земля - це незабудована власність, вільні земельні ділянки та сільськогосподарські угіддя (ферми, сади, ранчо та лісові масиви).
5. Нерухомість спеціального призначення - це нерухомість, що перебуває у загальному користуванні, наприклад, кладовища, урядові будівлі, бібліотеки, парки, школи [2].

Загальна економічна діяльність України має вплив як на ринок житлової нерухомості, так і на ринок комерційної нерухомості, оскільки вона визначає попит на нерухомість. Українська економіка розглядає нерухомість як один із

факторів виробництва, нарівні з працею та капіталом. Оптимальне використання нерухомості може суттєво вплинути на продуктивність бізнесу.

Згідно з практикою, існує прямий зв'язок між станом ринку нерухомості та макроекономічним розвитком країни. Рівень економічного зростання може значно впливати на попит на нерухомість. Наприклад, під час процвітання економіки та збільшення кількості робочих місць, попит на офісні приміщення зазвичай зростає. Зі збільшенням особистих доходів від нових робочих місць, люди мають більше коштів на покупки, що призводить до збільшення попиту на торгові приміщення. Крім того, коли працівники мають стабільний дохід, вони можуть дозволити собі власне житло, що сприяє попиту на ринку житла [3].

Ринок житлової нерухомості, де житло продається або придбається через посередників з нерухомості [4], відіграє невід'ємну роль у економіці країни. Цей ринок має значення не тільки на економічному рівні, але й на соціальному, оскільки він задовольняє одну з основних потреб людини – потребу в житлі [5, с. 19].

Ринок нерухомості залучає різні зацікавлені сторони, такі як власники нерухомості, які продають своє майно, особи, які бажають придбати нерухомість, орендарі, інвестори, які купують та продають нерухомість з метою інвестування, підрядники, які займаються ремонтом, і агенти з нерухомості, що діють як посередники в процесі купівлі або продажу нерухомості [6]. Згідно з Податковим кодексом України (пункт 14.1.129), об'єктами житлової нерухомості є будівлі, що належать до житлового фонду, а також дачні та садові будинки. Будівлі, що відносяться до житлового фонду, поділяються на наступні типи:

1. Житловий будинок - це сталі побудови, споруджені згідно з законодавчими та іншими нормативними актами і призначені для постійного проживання.

2. Прибудова до житлового будинку - це частина будинку, яка виходить за межі основних зовнішніх стін і має одну або більше спільних з основною частиною будівлі стін.

3. Квартира - самостійне помешкання з набором кімнат для постійного проживання, розташоване в житловому будинку.

4. Кімнати у багатоквартирних (комунальних) квартирах - окремі приміщення в квартирі, де проживають орендарі.

5. Котедж - невеликий будинок, зазвичай розташований за містом і призначений для постійного або тимчасового проживання, з приватною земельною ділянкою.

6. Садовий будинок - це будинок призначений для літнього (сезонного) використання, який не відповідає нормам, встановленим для житлових будинків, щодо забудованої площі, зовнішніх конструкцій та технічного обладнання. Дачний будинок - це житловий будинок, який можна використовувати протягом усього року для відпочинку на природі [7].

Житло може бути поділене на міське та заміське, а також на різні класи, такі як економ, комфорт, бізнес і преміум. Ринок житлової нерухомості можна розділити на дві категорії: первинний і вторинний. Первинний ринок нерухомості включає угоди, які здійснюються з новобудовами, тоді як вторинний ринок охоплює угоди з об'єктами, що перебувають у чиїйсь власності.

Зміни на ринку житла завжди мають важливе значення через взаємозв'язок між цінами на житло та споживчими витратами. Коли ціни на житло зростають, власникам власних будинків стає краще, оскільки вартість їхніх будинків збільшується. Цей "ефект багатства" підвищує довіру власників будинків і сприяє збільшенню споживання. Деякі люди можуть позичати більше під заставу свого будинку, витрачати на товари та послуги, ремонтувати свій будинок, доповнювати пенсію або погашати інші борги.

Проте, коли ціни на будинки знижуються, власники нерухомості можуть відчувати невпевненість, оскільки вартість їхнього майна зменшується. Це може призвести до скорочення споживання та відкладання особистих інвестицій від власників будинків. Крім того, купівля та продаж будинків безпосередньо впливають на економіку, крізь вплив на загальні витрати домогосподарств. Інвестиції в житло становлять невелику, але непередбачувану частину загального обсягу виробництва в економіці.

Існує два різних способи, якими купівля та продаж будинків впливають на ВВП. Перший спосіб полягає у придбанні житла, яке знаходиться на стадії будівництва. Це призводить до збільшення ВВП через інвестиції в землю для будівництва, закупівлю будівельних матеріалів та створення робочих місць. Коли власники будинків переїжджають, вони також сприяють місцевій економіці, здійснюючи покупки в місцевих магазинах.

Другий спосіб полягає у купівлі або продажу існуючого житла. Придбання існуючого будинку не має такого самого прямого впливу на ВВП, але все ж сприяє його зростанню. Наприклад, це включає витрати на послуги агентів з нерухомості, юридичні послуги, витрати на переїзд та покупку нових меблів.

Ціни на житлову нерухомість підлягають впливу закону попиту та пропозиції, аналогічно до інших активів, таких як акції та облігації. Попит на житло визначається як кількість нерухомості, яку покупці бажають та можуть придбати за певною ціною протягом певного періоду часу. Кілька факторів впливають на попит на житло:

1. Реальні доходи: Збільшення реальних доходів призводить до зростання попиту на житло, оскільки люди мають більше можливостей покращити свій рівень життя.

2. Вартість іпотечного кредиту: Зростання процентних ставок на іпотеку зазвичай призводить до зменшення попиту на житло, оскільки фінансування стає дорожчим, що робить купівлю нерухомості менш доступною.

3. Наявність кредиту: Чим більше банків та будівельних компаній готові надавати кредити, тим більше людей можуть позичати кошти на придбання житла, що збільшує попит та, відповідно, ціни.

4. Економічне зростання: У період економічного зростання та підвищення ділового циклу, заробітна плата зростає, що сприяє збільшенню попиту на житло.

5. Населення: Зростання населення або кількості домогосподарств призводить до збільшення попиту на житло.

6. Рівень зайнятості/безробіття: Чим більший в країні рівень безробіття, тим менша кількість людей зможе собі дозволити придбати нове житло.

7. Впевненість: Якщо споживачі мають оптимістичне ставлення до майбутнього економічного стану, то ймовірно, вони більш схильні до придбання житла, що призводить до зростання попиту. Ціни на житло нахиляються до підвищення, коли люди сподіваються на поліпшення свого фінансового стану у майбутньому.

Зниження цін на нерухомість призводить до збільшення кількості осіб, які можуть собі дозволити придбати житло, що призводить до збільшення попиту. Це є прикладом закону попиту, згідно з яким люди, коли ціна товару чи послуги знижується, стають більш схильними придбати більше благ.

Фактори, які впливають на пропозицію житла, можна умовно поділити на наступні категорії:

1. Витрати на виробництво: Вартість будівництва впливає на кількість будинків, що будуються і вводяться на ринок. Високі витрати, пов'язані з оплатою праці, земельними витратами та будівельними матеріалами, зменшують пропозицію житла.

2. Урядова політика: Державна політика, якщо вона впливає на оподаткування або субсидії для нового житла, може впливати на пропозицію.

Збільшення податків або зменшення субсидій може призвести до зменшення нових будинків на ринку.

3. Кількість будівельних компаній: Кількість будівельних компаній також може впливати на пропозицію житла. Залежно від їх цілей і масштабу діяльності, більше будівельних компаній може сприяти збільшенню пропозиції.

4. Технології та інновації: Вдосконалення технологій та впровадження інновацій у будівельній галузі можуть зробити будівництво більш ефективним і доступним. Це може призвести до збільшення пропозиції житла.

5. Державні витрати на нове соціальне житло: Держава впливає на пропозицію житла шляхом спонсорування та збільшення обсягів витрат на будівництво нових соціальних житлових об'єктів.

Пропозиція нового житла у короткостроковій перспективі є недостатньо еластичною щодо ціни. Основна причина цього полягає в тривалості процесу будівництва нового будинку. Від планування до завершення проекту може пройти багато місяців. Виробництво будинку також залежить від наявності кваліфікованої робочої сили та певних будівельних матеріалів. Через недостатню еластичність пропозиції, будь-які зміни в попиті можуть значно вплинути на ціни. Рівновага на ринку досягається тоді, коли попит та пропозиція збігаються. Це означає, що кількість, яку покупці хочуть придбати, і кількість, яку продавці хочуть продати, ідеально збігаються. Коли попит на житло високий, а пропозиція обмежена, ціни на житло зазвичай зростають. У разі перевищення пропозиції над попитом, власники будинків можуть знизити ціни, оскільки попит на ринку менш активний [8].

1.2 Фактори впливу на вартість нерухомості

Як вже було зазначено раніше, ринок нерухомості відіграє важливу роль у розвитку економіки, оскільки він сприяє залученню інвестиційного капіталу, стимулює виробництво і розвиток, збільшує доходи державного бюджету і розширює ринки, сприяючи соціально-економічній стабільності [9, с. 35].

Існує низка факторів, які можуть підвищувати або знижувати ціни на нерухомість. Тому важливо пильнувати за причинами, які можуть спричинити зріст або падіння цін на нерухомість [10].

Умовно, фактори, які впливають на ціну нерухомості, можна класифікувати на макро- та мікрофактори. Макрофактори, які впливають на вартість нерухомості, включають наступні аспекти:

1. Економічні фактори

По-перше, економічні фактори, які охоплюють загальні показники економіки і мають вплив на ринок нерухомості [11]. Фактори, такі як рівень безробіття, заробітна плата та приріст населення, впливають на наявність грошей у населення та на його можливість придбати житло [12]. Період зростання економіки, високий рівень зайнятості та доходів сприяють збільшенню кількості осіб, які можуть собі дозволити придбати нерухомість, що впливає на зростання цін на нерухомість [13]. За умов збільшення безробіття та сповільнення зростання заробітної плати, менша кількість осіб може дозволити собі придбати житло або матиме обмежені можливості переїзду [14]. Економічні фактори охоплюють такі показники, як рівень економічного розвитку країни, темпи зростання валового внутрішнього продукту, інфляційний рівень, тощо [15, с. 96].

2. Процентні ставки

Процентні ставки грають важливу роль у вартості житлових позик. Вони впливають на те, скільки покупців готові заплатити за нерухомість. Якщо ставка

нижча, покупці можуть собі дозволити придбати більш вартісний будинок. З іншого боку, вища ставка призводить до більших виплат за іпотеку, що зменшує доступну для них ціну придбання. Зі зростанням процентних ставок зменшується кількість покупців, які можуть собі дозволити купити будинок, оскільки оренда стає вигіднішою з плином часу.

3. Інвестиційний потенціал

Інвестиційний потенціал також впливає на вартість нерухомості для інвесторів. Фактори, такі як дохід від оренди та очікуваний капітальний приріст при подальшій продажу, відіграють свою роль.

4. Ринкові фактори

Ринкові фактори також мають важливе значення. Ринок нерухомості є одним з ключових ринків національної економічної системи. Водночас, він має пряму взаємозв'язок з іншими ринками вихідних ресурсів, такими як ринок праці та фінансовий ринок [9, с. 36].

Дослідження сучасних тенденцій на ринку також допомагає прогнозувати майбутню вартість будинків. Хоча історично вартість нерухомості має тенденцію до зростання, відомо, що вона може коливатися або падати у короткостроковій перспективі. Коли ціни на нерухомість починають знижуватися, багато покупців чекають, спостерігаючи за тенденціями зниження, що призводить до зменшення кількості активних покупців. Це, в свою чергу, призводить до збільшення пропонованих об'єктів та зниження вартості житла [17].

Ціни на будь-якому ринку визначаються попитом і пропозицією. Якщо високий попит на будинки у певній області, можна очікувати зростання їх вартості. З іншого боку, якщо на певному районі є зайвий запас нерухомості, але недостатньо покупців, то зазвичай ціни знижуються [12].

5. Населення і демографія

Фактори населення та демографії відображаються у соціальних характеристиках населення, включаючи житлові умови, потребу у земельних ділянках та нерухомості. Демографічні фактори включають тенденції у розмірі населення, розмірі сімей, рівні смертності та народжуваності тощо. Ці демографічні та соціальні фактори надають інформацію про попит на нерухомість та її розмір [15, с. 97].

6. Законодавчі фактори

Законодавчі фактори також мають значний вплив на попит на нерухомість та ціни. Податкові пільги, відрахування та субсидії є деякими засобами, якими уряд може тимчасово збільшити попит на нерухомість, поки вони існують [17]. До факторів державного регулювання входять такі елементи, як обмеження у продажі нерухомості, контроль над використанням землі, встановлення нормативних ставок орендної плати, регулювання будівельних норм, зонування та інші подібні заходи [15, с. 97].

7. Політичні фактори

Політичні фактори та стабільність також мають великий вплив на ринок нерухомості. Відсутність політичної єдності, революції, війни або конфлікти негативно відображаються на ринку нерухомості. Люди бояться можливих конфліктів, краху національної валюти та економічних криз, які можуть заважати продажу нерухомості або призводити до її знецінення [18].

Крім цього, на вартість рухомості також впливають політико-психологічні фактори, які включають:

- Рівень довіри громадян до банків та фінансових систем, ринку і майбутніх можливостей розвитку населених пунктів.
- Рівень оптимізму серед населення, який залежить від багатьох чинників, від перспективи розвитку країни до стану житлового фонду.
- Політична стабільність, що відображається у позитивних або негативних очікуваннях щодо ринку нерухомості [19, с. 252].

Цінним є оцінювання всіх цих макрофакторів, які мають вплив на вартість нерухомості. Проте, якщо потрібна більш детальна оцінка, розглядання мікрофакторів, які впливають на ціни на місцевому рівні, може бути корисним.

1. Розташування

Вартість нерухомості залежить від її розташування. Наприклад, нерухомість, яка знаходиться в центрі великого та густонаселеного міста, має вищу вартість порівняно з об'єктами, розташованими на околиці [9, с. 36]. Наприклад, чим ближче до межі Києва знаходиться заміський будинок, тим вища його ціна. Фактори, такі як близькість до багатих районів, торгових центрів або пляжів, можуть також збільшити вартість об'єкта. Зворотно, проживання біля аеропорту або в низько соціально-економічному районі може знизити ціну на нерухомість [12]. Крім того, наявність і якість інфраструктури поряд, таких як дитячі садки, школи, можливості працевлаштування, магазини, торгові, розважальні та рекреаційні центри, а також близькість до автомагістралей, комунікацій та громадського транспорту, можуть також впливати на загальну вартість будинку.

2. Планування булинку та розмір

Фактори, такі як розмір будинку та його планування, також мають вплив на його вартість. Вартість будинку зазвичай оцінюється в залежності від ціни за квадратний фут. Особливу увагу приділяють простору для проживання, який є найважливішим для покупців та оцінювачів. Спальні та ванні кімнати цінуються найбільше, тому наявність більшої кількості спалень та ванних кімнат зазвичай підвищує вартість будинку. Однак ці тенденції є дуже місцевими [13]. Наприклад, будинок з чотирма спальнями буде коштувати більше, ніж будинок з двома спальнями в тій самій області. Крім того, наявність додаткових ванних кімнат, гаражів, басейнів, дворів, відкритих розважальних зон і так далі, також має важливе значення для вартості нерухомості [11].

3. Вік та поточний стан майна

Вік та стан майна також впливають на його вартість. Зазвичай новіші будинки оцінюються вище. Це пояснюється тим, що важливі компоненти будинку, такі як сантехніка, електрика, дах та побутова техніка, зазвичай є новішими і менш схильними до поломок, що може бути вигідним для покупця [13]. Проведення регулярного технічного обслуговування є одним із ключових факторів, які можуть знизити вартість будинку. Також будь-які оновлення та модернізації, такі як установка нової техніки на кухні або реконструкція ванної кімнати, покращують стан будинку і впливають на його вартість [16].

4. Потенціал ремонту та покращень

Якщо покупець може вдосконалити майно, особливо в старих будинках, це може призвести до збільшення його вартості [15]. Можливість розширення є важливою як для покупців нерухомості, так і для інвесторів, оскільки вони можуть додати додаткову спальню, збільшити площу приміщення або встановити басейн або внутрішній дворик. Якщо покупець має можливість покращити та персоналізувати майно шляхом ремонту будинку, це також призведе до зростання його вартості при перепродажу.

5. Вплив місцевого ринку

Навіть якщо будинок перебуває у відмінному стані, знаходиться у найкращому місці та має преміальні оновлення, кількість інших нерухомостей, що продаються у цьому районі, а також кількість покупців на ринку можуть вплинути на його вартість. Якщо в сусідніх об'єктах нерухомості попит невеликий, а їх ціни значно нижчі від ринкової вартості, можна очікувати зниження значення будинку [11].

1.3 Тенденції розвитку ринку нерухомості в Україні

Протягом першої половини 2021 року в Україні спостерігався рекордний пік нового житлового пропозиції. Було введено в експлуатацію 11,4 млн квадратних метрів житла, що є найвищим показником за останні 30 років.

У 2022 році початок був також перспективним і міг бути продовжений у тому ж дусі. До початку війни темпи введення нового житла в експлуатацію залишалися на рівні 2021 року. За перші півтора місяці загальна площа прийнятих в експлуатацію житлових будівель становила майже п'яту частину очікуваного щорічного обсягу.

Після повномасштабного вторгнення практично всі будівельні компанії призупинили роботу на будівельних майданчиках: деякі - на кілька тижнів, а деякі до сих пір не поновили будівельні роботи. В результаті, у 2022 році загальна площа введеного в експлуатацію житла скоротилася до 7,1 млн квадратних метрів (18 300 приватних будинків і 74 300 квартир), що на 38% менше, ніж у 2021 році.

Зменшення обсягів нової пропозиції стало характерним практично для всієї території України, за винятком кількох областей. За даними онлайн-сервісу продажу нерухомості ЛУН, найбільший спад в будівництві спостерігався в східних та південних областях, де обсяги введення нового житла знизилися на 70-90% порівняно з попереднім роком [20].

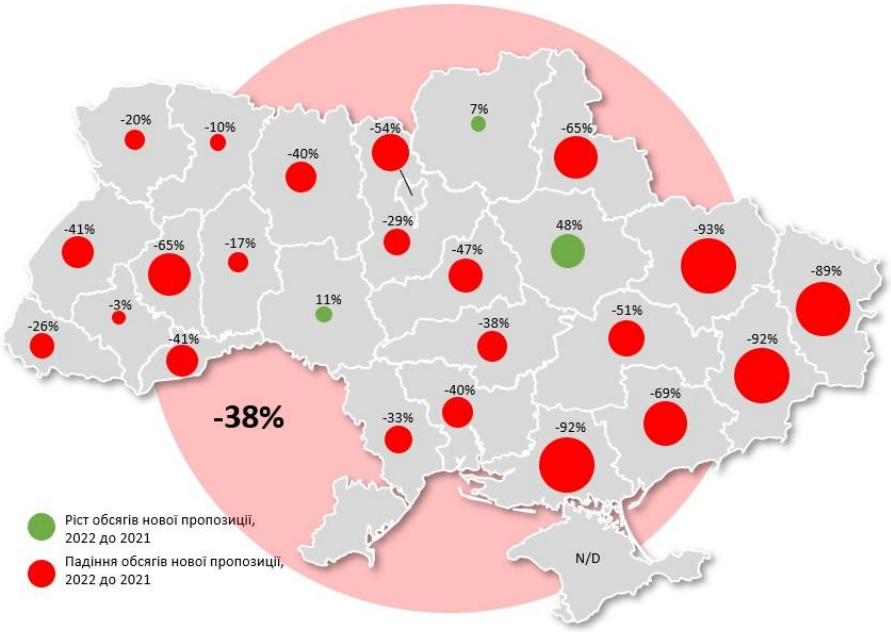


Рисунок 1.1 - Динаміка обсягів нової пропозиції на ринку житлової нерухомості України, 2022-й до 2021 року [22]

У зв'язку з новими житловими проєктами, передбачається зменшення їх кількості. Будівництво на нових будівельних майданчиках розпочалося лише для житлових комплексів, чиї проекти вже були затверджені, готові до запуску та мають початкове фінансування. Таким чином, протягом перших трьох кварталів 2022 року було розпочато будівництво лише 68 200 квартир, що вдвічі менше, ніж за аналогічний період попереднього року. [21]

Обсяги будівництва в Україні сповільнилися не лише через ризики фізичної безпеки, але й через значне зростання курсу валют і високу інфляцію, що зменшили купівельну спроможність населення, а також збільшили вартість будівництва.

Це призвело до нестачі коштів у більшості девелоперських проєктів, особливо на початковій стадії, що унеможливлювало продовження будівництва. Більше того, через зруйнування промислових підприємств на ринку виникла нестача будівельних матеріалів та проблеми з їх логістикою.

Проте, як повідомляють багато девелоперів, ці питання вже вирішені, оскільки забудовникам вдалося адаптуватися до нових умов і налагодити нові логістичні ланцюги з українськими та іноземними постачальниками.

З жовтня до грудня 2022 року будівельні компанії зіткнулися з новим викликом, а саме потребою відповідати графікам відключень електроенергії або придбати потужні генератори для безперебійного будівництва. Деякі забудовники вирішили призупинити будівництво, що також призвело до зниження обсягів.

Артилерійські та ракетні обстріли стали найскладнішою проблемою на ринку житлової нерухомості багатьох регіонів. За даними, активні бойові дії до вересня 2022 року призвели до зруйнування або пошкодження 74,1 млн квадратних метрів житла, що становить понад 7% від загального житлового фонду.

Розміри руйнувань є величезними, перевищуючи обсяги всього нового житла, зведеного за останні сім років, за наявними даними. Ураховуючи масштаби ракетних обстрілів в останні місяці 2022 року та неможливість оцінки збитків у окупованих районах, обсяги руйнування житла стануть ще більшими.

Війна спонукала людей більш обережно ставитися до покупок, особливо таких значних, як придбання житла. У перших місяцях конфлікту не відбувалося транзакцій у секторі житлової нерухомості, в основному через відсутність доступу до Державного реєстру речових прав на нерухоме майно. З травня 2022 року реєстри почали поступово відкриватися, що призвело до повільного відновлення попиту.

Проте, через невизначеність та фізичні та економічні ризики більшість потенційних покупців та інвесторів прийняла вичікувальну позицію, незалежно від регіону.

За словами забудовників, попит на квартири на первинному ринку становить лише 30% від попереднього рівня в західних областях, тоді як у Києві загальний попит становить 10-20% від попереднього рівня.

На другорядному ринку спостерігаються схожі тенденції: кількість угод купівлі-продажу квартир та житлових будинків в Україні у 2022 році склала 101 000 транзакцій, що становить лише третину обсягів попереднього року. Аналогічна ситуація спостерігається у більшості регіонів країни, з меншим зниженням активності (-40-60%) у безпечніших областях Центральної та Західної України, та найбільшим зменшенням попиту (-70-90%) на Сході та Півдні.

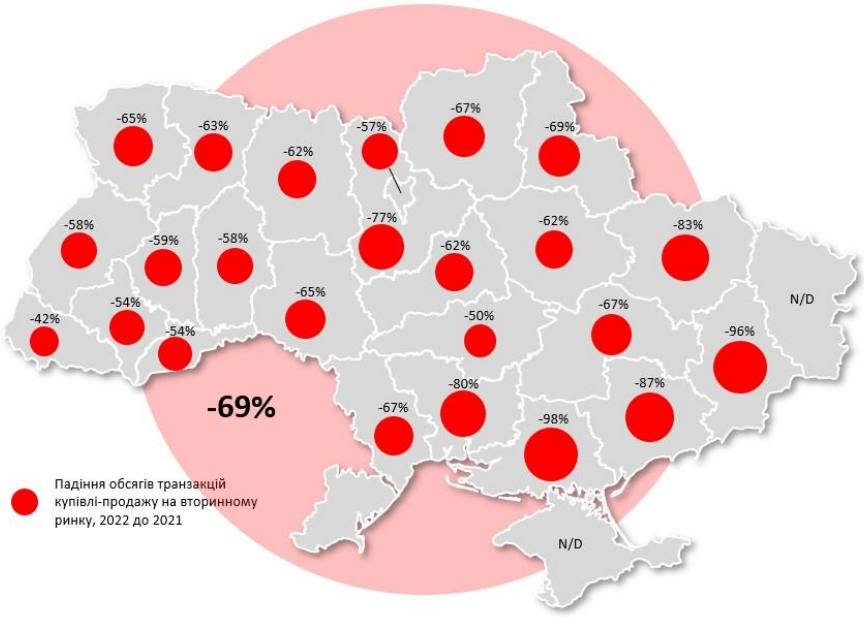


Рисунок 1.2 - Динаміка кількості укладених угод купівлі-продажу житла на вторинному ринку України, 2022-й до 2021 року [22]

Починаючи з жовтня 2022 року, в Україні була впроваджена державна програма під назвою "єОселя", яка має на меті забезпечити доступне кредитування. Поки що ця програма доступна тільки для певних соціальних категорій населення, таких як військові, правоохоронці, медики та працівники

освіти. Ці особи можуть отримати кредит на житло з пільговою ставкою 3% річних на строк до 20 років.

На початку 2023 року планувалося розширити список громадян, які можуть отримати кредит на житло під 7%. Однак через воєнний стан і недостатнє фінансування, запуск масової іпотеки був відкладений. За словами керівників проєкту, очікується, що програма буде запущена протягом 2023 року.

За даними компанії "Укрфінжитло", яка є координатором проєкту "єОселя", станом на 31 грудня 2022 року (після трьох місяців роботи) 451 сім'я придбала власне житло за допомогою цієї програми. Більшість з них складають сім'ї військовослужбовців та правоохоронців.

Але, починаючи з 31 січня, компанія "Укрфінжитло" оголосила, що кошти з першого етапу фінансування програми були повністю розподілені та використані банками-учасниками. Це означає, що навіть тим, хто був затверджений (близько 10 000 кандидатів), фінансові установи поки що не видають кошти. Організатори заявляють, що наступний етап кредитування розпочнеться вже цієї весни після залучення додаткових ресурсів. [23]

Український ринок нерухомості, що історично формувався в іноземній валюті, є одним із секторів економіки. Через це українські інвестори часто вкладають свої кошти в нерухомість, вважаючи це безризиковим.

На кінець 2022 року немає чіткої тенденції в динаміці цін на ринку купівлі-продажу житла. З одного боку, витрати на будівництво зросли через знищення деяких заводів-виробників будівельних матеріалів, ускладнення ланцюгів постачання сировини, девальвацію гривні та додаткові витрати на забезпечення роботи під час відключень електроенергії. З іншого боку, обмежений попит не дозволяє забудовникам значно підвищити вартість продажу квартир.

Динаміка цін на продаж та оренду житла відрізнялась у різних регіонах країни. За даними онлайн-сервісу продажу нерухомості DOM.RIA, в більшості

західних і північно-західних регіонів України, а також в областях з великим потоком внутрішньо переміщених осіб, таких як Запорізька, Миколаївська, Дніпропетровська області, ціни на квартири всіх класів зростали або залишалися на приблизно тому ж рівні, що й раніше. На інших територіях країни житло в середньому стало дешевшим на 2-10% (з найбільшим зниженням до 20% у Харківській області).

На вторинному ринку ситуація стає ще більш складною, оскільки власники більш схильні зайняти очікувальну позицію, щоб уникнути втрати можливих доходів від продажу в умовах спаду ринку. Однак, якщо є реальний інтерес, власники часто йдуть на компроміси з малою кількістю потенційних покупців, щоб залучити їх.

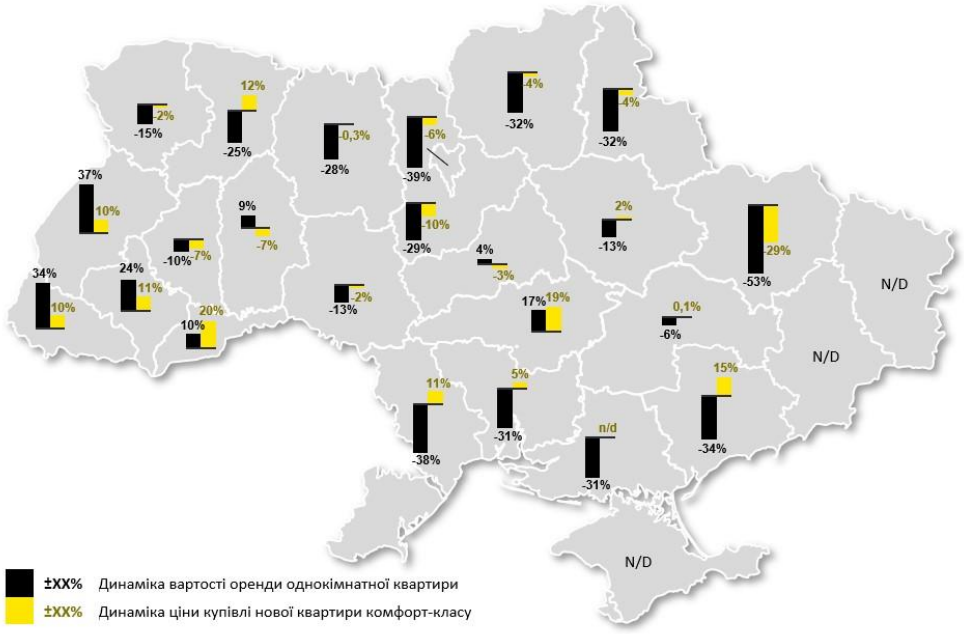


Рисунок 1.3 - Динаміка середніх цін (в доларах США) на оренду та купівлю квартир по областях, грудень 2022-го до грудня 2021 року [24]

В 2023 році передбачається, що забудовники будуть продовжувати спрямовувати всі свої ресурси на завершення проектів, які знаходяться на останній стадії будівництва. Наразі немає прогнозів щодо появи нових

будівельних проектів у найближчому майбутньому. У випадку продовження затримок у поточних будівельних проектах та відкладення нових проєктів у середньостроковій перспективі, на ринку може виникнути нестача пропозиції житлових площ.

Одночасно, що стосується попиту, незважаючи на низьку покупчу спроможність та загальну невизначеність на ринку, в найближчому майбутньому не передбачається значних позитивних змін. Після завершення війни та стабілізації економіки очікується поступове відновлення покупчої активності. Тоді можливо, що відкладений попит на ринку нерухомості проявиться. Також, особи, які були внутрішньо переміщені і залишаться на своїх поточних місцях проживання, ймовірно, розглядатимуть можливість придбання власного житла. У довгостроковій перспективі відновлення попиту буде залежати від темпів макроекономічної стабілізації після завершення війни.

РОЗДІЛ 2. АНАЛІЗ МЕТОДІВ ТА МЕТОДОЛОГІЙ ІНФОРМАЦІЙНОГО АНАЛІЗУ ТА МОДЕЛЮВАННЯ

2.1 Огляд та опис методології побудови лінійної регресійної моделі

Статистичне моделювання включає в себе процес застосування статистичного аналізу до набору даних. Статистична модель визначає математичну залежність між однією або кількома випадковими величинами та іншими невипадковими величинами [25].

У контексті контрольованих методів навчання використовуються моделі регресії і моделі класифікації [26]. Регресійний аналіз у статистичному моделюванні використовується для виявлення залежності між однією або кількома залежними змінними (зазвичай позначаються як Y) і незалежними змінними (які зазвичай позначаються як x1, x2, ..., xp). Цілями регресійного аналізу є:

1. Встановлення випадкового зв'язку між змінною відповіді Y та регресорами x1, x2, ..., xp.
2. Прогнозування значень Y на основі набору значень x1, x2, ..., xp.
3. Визначення, які змінні є найважливішими для пояснення змінної відповіді Y, з метою більш ефективного і точного визначення причинно-наслідкового зв'язку [27, с. 4].

Регресійний аналіз відрізняється від класифікаційних моделей тим, що він оцінює числові значення, в той час як класифікаційні моделі визначають категорію, до якої належить спостереження [28]. У емпіричних науках регресійний аналіз є одним з найчастіше використовуваних математичних методів [29, с. 4]. Основні напрямки регресійного аналізу включають прогнозування, моделювання часових рядів і пошук причинно-наслідкових зв'язків між змінними [28].

Аналітики часто використовують, але не обмежуються такими процедурами при регресійному аналізі:

1. Розуміння конкретної проблеми, пов'язаної з певною науковою галуззю;

2. Необхідність визначення математичної моделі регресії, яку можна записати за допомогою формули (2.1).

$$Y = f\ (x1, x2, \dots, xp)\ + \varepsilon, \qquad (2.1)$$

3. Провести статистичні перевірки, щоб зробити припущення про розподіл випадкової помилки ε у моделі регресії. Ці припущення потрібно перевірити за допомогою даних, які зібрані в результаті експерименту.

4. Зібрати дані Y та x1, x2, ..., xp. Цей крок зазвичай включає експериментальне проектування, визначення обсягу вибірки, створення бази даних, очищення даних та виведення змінних, які будуть використовуватися для статистичного аналізу.

5. Створити набори даних у відповідному форматі відповідно до використовуваного програмного забезпечення для аналізу.

6. Уважно оцінити, чи підходить обрана модель для відповіді на поставлені наукові запитання.

7. Якщо модель задовольняє загальновизнаним критеріям діагностики моделі, її можна використовувати для відповіді на бажані наукові питання. В іншому випадку модель може бути вдосконалена або модифікована [27, с. 4-5].

Існує багато методів регресійного аналізу, і вибір конкретного методу залежить від різних факторів, таких як тип цільової змінної, форма лінії регресії та кількість незалежних змінних [30].

Лінійна регресія є одним з найбільш відомих та зрозумілих алгоритмів статистики та машинного навчання. Вона представляє собою лінійну модель, яка передбачає лінійну залежність між вхідними змінними (X) та єдиним вихідним параметром (У). Якщо вхідна змінна (X) є одна, то метод називається

простою лінійною регресією. У випадку, коли є кілька вхідних змінних, цей метод позначається як множинна лінійна регресія [31] у статистичній літературі.

Проста лінійна регресія передбачає підхід до прогнозування кількісного відгуку залежної змінної (У) на основі єдиної незалежної змінної (Х). Цей підхід базується на припущенні про існування значущого лінійного зв'язку між Х та У. Математично, просту лінійну регресію можна записати за допомогою формули (2.2):

$$Y \approx \beta_0 + \beta_1 X + \varepsilon, \tag{2.2}$$

де $\beta 0$ – вільний член рівняння, тобто очікуване значення У при Х=0;

$\beta 1$ – кут нахилу лінійної моделі;

ε – помилка, яка не залежить від Х та має нульове середнє значення.

Після того, як модель була уточнена і дані були зібрані, наступним кроком є пошук оцінок β_0 та β_1 для простої лінійної регресійної моделі, яка найкраще відповідає отриманим даним після проведення наукового експерименту [27, с. 9]. Існує кілька способів знаходження параметрів цієї моделі, але найпоширеніший підхід полягає у мінімізації критерія найменших квадратів. Метод найменших квадратів є статистичною процедурою для знаходження найкращої прямої, яка найбільш точно пасує до набору даних, шляхом мінімізації суми квадратів відхилень між точками та побудованою прямою [32]. Після проведення певних обчислень можна показати, що формула (2.3) використовується для визначення оцінок коефіцієнтів простої лінійної регресії з використанням методу найменших квадратів.

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(yi - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \widehat{\beta_0} = \bar{y} - \beta_1 \cdot \bar{x} \tag{2.3}$$

де $\bar{y}=$ та $\bar{x}$– вибіркові середні.

Ми можемо використати наші навчальні дані, щоб знайти оцінки для коефіцієнтів β_0 та β_1. За допомогою формули (2.4) [33, с. 73], ми можемо прогнозувати значення Y на основі нових значень X.

$$\hat{Y} \approx \widehat{\beta_0} + \widehat{\beta_1} X + \varepsilon, \tag{2.4}$$

де $\hat{Y}$- це значення Y, передбачене по X.

В лінійній регресії велику роль відіграє поняття стандартної помилки. Стандартна помилка, яка є приблизним стандартним відхиленням статистичної вибіркової сукупності, має важливе значення. Вона вказує на те, наскільки точними є прогнози значень змінної Y на основі значень X. Чим ближче спостереження на діаграмі розсіювання до лінії регресії, тим менша стандартна помилка [34]. Щоб визначити стандартні похибки коефіцієнтів регресії, можна скористатися формулами (2.5) [33, с. 78].

$$SE\left(\widehat{\beta_0}\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$

$$SE\left(\widehat{\beta_1}\right)^2 = \sigma^2 \left[\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \tag{2.5}$$

Для перевірки статистичних гіпотез про відношення коефіцієнтів, можна скористатися стандартними помилками [35]. Найпоширенішим є метод перевірки нульової гіпотези - H_0: взаємозв'язку між X та Y немає, проти альтернативної гіпотези - H_a: між X та Y існує певний зв'язок. Для перевірки нульової гіпотези необхідно визначити, наскільки далеко від нуля знаходиться оцінка β_1, щоб бути впевненими у її значущості. У практиці застосовується t-критерій Стьюдента для всіх параметрів моделі, який показує їх статистичну значущість. Значущість параметрів вказує на ймовірність їх відмінності від нуля. В даному випадку нульова гіпотеза стверджує, що $\beta_1 = 0$. З формули (2.6) можна отримати значення t-критерія:

$$t_j = \frac{\beta_i - 0}{SE(\widehat{\beta_1})} \tag{2.6}$$

Кожне обчислене значення порівнюється з критичним значенням на встановленому рівні значущості та ступені свободи. Якщо обчислене значення перевищує критичне, то відповідний параметр моделі вважається значущим; навпаки, якщо обчислене значення менше критичного, параметр вважається нестатистично значущим. Іншими словами, припустивши, що $\beta_1 = 0$, легко обчислити ймовірність будь-якого значення, яке дорівнює або перевищує $|t_j|$. Ця ймовірність називається p-значенням. P-значення інтерпретується так: якщо мале p-значення підтверджує наявність реального зв'язку між предиктором та відповіддю, це означає низьку ймовірність випадкового виявлення відсутності зв'язку між цими змінними [33, с. 79]. Рівень статистичної значущості часто виражається як p-значення від 0 до 1. Чим менше p-значення, тим сильніші докази на користь відхилення нульової гіпотези.

– значення $p \leq 0{,}05$ вказує на статистичну значимість. Це означає, що існує вагомі докази проти нульової гіпотези, оскільки ймовірність відсутності залежності між змінними менше 5%. Тому ми відхиляємо нульову гіпотезу і приймаємо альтернативну гіпотезу.

– значення $p > 0{,}05$ не є статистично значущим і свідчить вагомі докази на користь нульової гіпотези. Це означає, що ми залишаємо нульову гіпотезу і відхиляємо альтернативну гіпотезу [36].

Після відхилення нульової гіпотези на користь альтернативної можна оцінити, наскільки добре модель оцінює дані. Якість лінійної регресійної моделі оцінюється за допомогою різних показників. Дві найбільш часто використовувані міри, що залежать від масштабу, базуються на абсолютних помилках або помилках у квадраті: MSE та RMSE. У статистиці середньоквадратична помилка MSE вимірює середнє значення квадратів помилок, тобто середня квадратична різниця між розрахунковими значеннями та фактичними величинами, що можна побачити у формулі (2.7) [37].

$$MSE = \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{n}, \quad (2.7)$$

де $\hat{y}$- це значення Y, передбачене по X.

Квадратний корінь цього значення позначається як RMSE і обчислюється за формулою (2.9):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{n}} \quad (2.8)$$

Відсоткові помилки мають певну перевагу, оскільки вони позбавлені одиниць виміру, і тому їх часто використовують для порівняння прогнозованих показників між різними наборами даних. Два найбільш поширених показники цього типу - це MAPE і sMAPE. MAPE - це середня абсолютна відсоткова помилка, що обчислюється за формулою (2.9):

$$MAPE = \frac{1}{n} * \sum_{i=1}^{n} \frac{|yi-\hat{y}i|}{yi} * 100\% \quad (2.9)$$

sMAPE – симетрична середня абсолютна відсоткова помилка, яка розраховується за формулою (2.10):

$$sMAPE = \frac{100\%}{n} * \sum_{i=1}^{n} \frac{|yi-\hat{y}i|}{\frac{(|yi|+|\hat{y}i|)}{2}} \quad (2.10)$$

Часто використовуються коефіцієнти множинної детермінації (R^2) і скоригований коефіцієнт множинної детермінації (R^2_{adj}) для оцінки відповідності моделі реальним значенням даних. Коефіцієнт R^2 вказує, наскільки точки даних відповідають кривій або лінії. Скоригований коефіцієнт R^2 також показує, наскільки точки даних відповідають кривій або лінії, але враховує кількість членів у моделі. Коефіцієнти обчислюються за формулами (2.11) та (2.12).

$$R^2 = 1 - \frac{D(\hat{y})}{D(y)}, \quad (2.11)$$

де $D(\hat{y})$ – детермінована дисперсія; $D(y)$ – загальна дисперсія.

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}, \quad (2.12)$$

де n – кількість спостережень у наборі даних; k– кількість незалежних регресорів.

Формула (2.13) використовується для обчислення коефіцієнта кореляції (R), який відображає ступінь зв'язку між залежною та незалежною змінною.

$$R = \pm\sqrt{R^2} \tag{2.13}$$

Основною ціллю множинної лінійної регресії є виявлення лінійної залежності між залежною змінною та кількома незалежними змінними. Загальний вигляд моделі множинної лінійної регресії визначається формулою (2.14).

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_k X_k + \varepsilon, \tag{2.14}$$

де β_0 – вільний член рівняння;

β_k – коефіцієнт, який виражає силу зв’язку між змінною та відгуком;

X_k – k-й предиктор;

ε – помилка, яка не залежить від X та має нульове середнє значення.

Ми розглядаємо β_k як середню зміну Y, викликану збільшенням β_k на одну диницю при утриманні всіх інших предикторів на одному рівні. Аналогічно до простої лінійної регресії, параметри оцінюються за допомогою методу найменших квадратів. Ми обираємо β_0, β_1,..., β_k з метою мінімізації суми квадратів залишків.

Прогнозування Y також можна здійснити за тим самим принципом, що і в простій лінійній регресії, використовуючи формулу (2.15).

$$Y \approx \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + .. + \widehat{\beta_k} X_k + \varepsilon, \tag{2.15}$$

Зазвичай, коли ми працюємо з множинною лінійною регресією, нас цікавлять деякі важливі питання. Перше питання - чи існує зв'язок між залежною та незалежними змінними? Щоб встановити наявність такого зв'язку, ми повинні перевірити статистичну гіпотезу $H_0 : \beta_1 = \beta_2 = ... = \beta k = 0$ проти альтернативної гіпотези H_a: принаймні один βk не дорівнює нулю [33, с. 79-85]. Для перевірки цієї гіпотези використовується F-критерій, який відображає,

наскільки добре модель пояснює загальну дисперсію залежної змінної. Його значення обчислюється за формулою (2.16).

$$F = \frac{R^2}{1-R^2} \cdot \frac{f_2}{f^1}, \quad (2.16)$$

Щоб перевірити значущість рівняння регресії, розраховане значення тесту Фішера порівнюють з табличним значенням на обраному рівні значущості (зазвичай 0,05), яке відповідає числу ступенів свободи f1 (більша дисперсія) та f2 (менша дисперсія) [38]. Якщо значення цієї статистики перевищує критичне значення на даному рівні значущості, то нульова гіпотеза відхиляється, що означає статистичну значимість регресії. Якщо ж значення не перевищує критичне значення, то модель вважається незначущою. Також варто зазначити, що якщо між відповіддю та предикторами немає зв'язку, очікується, що значення F-критерію буде наближене до 1, а в разі справедливості альтернативної гіпотези очікується, що F > 0. За допомогою значення p, так само як і у простій лінійній регресії, можна зробити висновок про ймовірність прийняття нульової гіпотези.

Які методи можна використовувати для підбору важливих предикторів? Ці методи спрямовані на визначення набору змінних, які найкраще відповідають моделі, щоб забезпечити точні прогнози. Ця задача відома як відбір інформаційних змінних. Для розв'язання цієї задачі існує три класичних методи:

Відбір із включенням: Починаючи з нульової моделі, яка включає в себе лише вільний член, без предикторів, ми додаємо по одній змінній, яка призводить до найменшої суми квадратів залишків. Цей процес триває до задоволення певного правила зупинки.

Відбір із виключенням: Починаючи з моделі, яка включає всі змінні, поступово виключаємо змінні з найбільшим значенням p. Потім ми побудуємо нову модель з (p-1) змінними і виключимо з неї змінну з найбільшим p значенням. Цей процес триває до задоволення певного правила зупинки.

Комбінований відбір: Цей метод поєднує два попередніх підходи. Ми починаємо з нульової моделі і поступово додаємо змінні, щоб забезпечити найкращу якість моделі. За мірою додавання нових предикторів, деякі значення p можуть зростати. Якщо значення p перевищує певний поріг, ми виключаємо цю змінну з моделі. Цей процес продовжується до тих пір, поки значення p для всіх змінних в моделі не стане достатньо низьким, а значення p для виключених змінних буде достатньо високим для включення в модель.

Існують різні критерії, які можна використовувати для оцінки якості моделі. Серед них найбільш поширеними є Акаїкевський інформаційний критерій, Баєсівський інформаційний критерій та скорегований коефіцієнт детермінації R^2 [33, с. 87-95].

2.2 Опис методології побудови регуляризації

Однією з основних проблем, з якими стикаються фахівці у сфері обробки даних, є проблема перенавчання. Перенавчання виникає, коли модель надмірно аналізує шум та деталі в навчальних даних, що негативно впливає на її продуктивність при роботі з новими даними. Іншими словами, модель стає надто складною, так що помилка на навчальних даних зменшується, але помилка на тестових даних зростає [40, с.182]. Регуляризація - це набір методів, які використовуються для зменшення цієї помилки шляхом належного підбору функції до навчального набору та уникнення перенавчання [41]. Як було зазначено у розділі 2.1 цієї роботи, просте рівняння для лінійної регресії має наступний вигляд [див. формулу (2.17)], де Y є залежним змінним, а β_k - оцінками коефіцієнтів для різних змінних або предикторів (X).

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_k X_k \qquad (2.17)$$

Коефіцієнти підбираються таким чином, щоб мінімізувати функцію втрат, відому як залишкова сума квадратів або RSS, яку можна бачити з формули (2.18), в процедурі підгонки.

$$RSS = \sum_{i=1}^{n}\left(yi - \beta_0 - \sum_j^p \beta_j x_{ij}\right)^2 \quad (2.18)$$

Якщо в навчальних даних присутні шуми, то отримані коефіцієнти не зможуть добре узагальнюватись на майбутні дані. Це проблему можна вирішити за допомогою регуляризації, яка зменшує або обмежує ці оцінки до нуля. Іншими словами, регуляризація є формою регресії, що контролює або зменшує значення оцінок коефіцієнтів [42]. Два найпоширеніших типи регуляризації - L1 (Lasso) і L2 (Ridge). Вони впливають на загальну функцію втрат, додавши до неї додатковий термін, відомий як термін регуляризації [43]. Формула (2.19) описує регресію типу Ridge, де RSS модифікується додаванням певної кількості стиснення. Коефіцієнти оцінюються шляхом мінімізації цієї функції [42].

$$\sum_{i=1}^{n}\left(yi - \beta_0 - \sum_j^p \beta_j x_{ij}\right)^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \cdot \sum_{j=1}^{p} \beta_j^2, \quad (2.19)$$

де λ - є параметром штрафування.

Рідж регресія, так само, як метод найменших квадратів, прагне знайти оцінки коефіцієнтів, що забезпечують ефективний опис даних шляхом мінімізації суми квадратів помилок (RSS). Але другий член, який називається штрафним додатком, набуває невеликих значень, коли коефіцієнти β_1 , ..., β□ наближаються до нуля, що сприяє стисненню оцінок β в напрямку нуля [33, с. 234]. Збільшення гнучкості моделі передбачає збільшення її коефіцієнтів, а якщо ми хочемо мінімізувати вищезазначену функцію, то ці коефіцієнти мають бути невеликими.

Отже, регуляризація рідж регресії дозволяє уникнути надмірно великих коефіцієнтів [42]. Гіперпараметр λ є налаштувальним параметром, який визначає, наскільки ми бажаємо штрафувати гнучкість нашої моделі [43].

Якщо λ = 0, то штрафний доданок не має жодного ефекту, і рідж регресія дасть ті самі оцінки, що й метод найменших квадратів. Проте, коли $\lambda \to \infty$, вплив штрафного доданку зростає, і оцінки коефіцієнтів рідж регресії наближатимуться до нуля [33, с. 235]. Тому вибір правильного значення λ є

критичним. При оцінці різних параметрів, таких як λ, все ще існує ризик перенавчання тестового набору, оскільки параметри можна налаштовувати до тих пір, поки оцінювач не працюватиме оптимально. Таким чином, знання про тестовий набір може "просочитися" в модель, і показники оцінки більше не будуть відображати результатів узагальнення.

Для вирішення даної проблеми застосовується процедура, відома як перехресна перевірка. У базовому підході, що називається k-кратною перехресною перевіркою, навчальний набір ділиться на k менших наборів (зазвичай 5-10). Для кожного з наборів виконується така послідовність дій:

1. Модель навчається за допомогою k-1 набору даних як тренувальних даних.

2. Отримана модель перевіряється на решті даних (використовується як тестовий набір) для оцінки її ефективності, наприклад, точності.

3. Записуються показники ефективності для кожного набору, а потім обчислюється їх середнє значення. Це буде показник ефективності моделі [44].

Таким чином, при використанні регуляризації перехресна перевірка дозволяє знайти оптимальне значення параметру λ, яке призводить до найнижчої помилки перехресної перевірки (значення MSE) [42].

Слід зазначити, що регуляризація Ridge просто стискає коефіцієнти до нуля, але не робить їх рівними нулю. Це може не впливати на точність прогнозування, але ускладнює інтерпретацію моделі, особливо коли кількість предикторів є великою. Метод Lasso є альтернативою регуляризації Ridge і може вирішити цей недолік.

Метод Lasso мінімізує коефіцієнти за допомогою формули (2.20) [33, с. 239]:

$$\sum_{i=1}^{n}\left(yi - \beta_0 - \sum_{j}^{p}\beta_j x_{ij}\right)^2 + \lambda \cdot \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \cdot \sum_{j=1}^{p}|\beta_j| \qquad (2.20)$$

Зрозуміло, що ця варіація відрізняється від рідж регресії тим, що вона використовує покарання високих коефіцієнтів, представлене модулем $|\beta_j|$,

замість квадратів β, як у рідж регресії. Це підходить до статистичного поняття норми L1 [42]. Ласо, схоже на рідж регресію, стискає оцінки коефіцієнтів у напрямку нуля. Проте, в разі ласо, l1-штраф при достатньо великому значенні параметра λ, приводить деякі коефіцієнти до точного нульового значення. Іншими словами, можна сказати, що ласо здійснює відбір змінних. Як і в рідж регресії, вибір значення параметра λ є критичним, і його також можна визначити за допомогою перехресної перевірки. Загалом, можна сказати, що ласо регресія працює краще, коли в моделі є багато непотрібних предикторів, оскільки їх можна легко виключити; з іншого боку, рідж регресія працює краще, коли більшість предикторів є потрібними. Однак часто в моделі існує велика кількість предикторів, і важко визначити, чи вони є потрібними. Щоб уникнути вибору між цими двома методами, можна скористатися методом еластичної сітки, який комбінує їх. Він використовує як рідж, так і ласо покарання для регуляризації. Регресія еластичної сітки групує та зменшує параметри, пов'язані з корельованими змінними, залишаючи їх у рівнянні або видаляючи відразу [див. формулу (2.21)].

$$RSS + \lambda \cdot \left(\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right) \qquad (2.21)$$

Крім встановлення та вибору значення λ, еластична сітка також дозволяє налаштувати параметр альфа. Значення альфа від 0 до 1 визначає тип регресії: 0 відповідає рідж регресії, а 1 - лассо регресії. Це дозволяє оптимізувати еластичну сітку [45].

Варто зауважити, що звичайні оцінки коефіцієнтів, отримані методом найменших квадратів, є еквіваріантними. Це означає, що множення всіх коефіцієнтів на одну і ту саму константу просто зменшує ці оцінки пропорційно. З іншими словами, шкала вимірювання не впливає на величину коефіцієнта β. З іншого боку, оцінки коефіцієнтів при регуляризації регресії можуть значно змінюватися при множенні певного предиктора на константу.

Тому перед застосуванням регуляризації необхідно стандартизувати або привести предиктори до одного масштабу. Для цього можна використати формулу (2.22) [33, с. 237].

$$\widetilde{x_{ij}} = x_{ij} / \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_{ij} - \bar{x_j}\right)^2} \qquad (2.22)$$

2.3 Опис методології побудови алгоритму Random Forest для машинного навчання

Перш ніж розглядати алгоритми машинного навчання, такі як "Randomforest" або "Випадковий ліс", слід спочатку розглянути концепцію дерев прийняття рішень.

Дерева рішень - це метод навчання, що не вимагає параметрів, використовуваний як для класифікаційних, так і для регресійних задач. Його мета полягає в створенні моделі, яка передбачає значення цільової змінної, вивчаючи прості правила прийняття рішень, отримані з особливостей даних [46]. Дерево рішень будує моделі регресії або класифікації у вигляді структури дерева. Воно розбиває набір даних на все менші підмножини, тоді як дерево рішень поступово розгалужується. Дерева рішень можуть обробляти як категоріальні, так і числові дані [47].

Зазвичай дерева рішень складаються з трьох основних елементів. Перший елемент - кореневий вузол - є верхнім рівнем дерева і представляє кінцеву мету або рішення, яке потрібно прийняти. Другий елемент - гілки - відображають різні варіанти або напрямки дій, доступні при прийнятті рішення. Третій елемент - вузли або листки - представляє можливі результати для кожної дії і розташовується в кінці гілок. Вузли можуть бути внутрішніми, що вказують на інші рішення, або листовими, які вказують на випадкову подію чи невідомий результат.

Для кращого розуміння процесу створення дерева рішень можна використовувати наступний алгоритм:

1. Почати з кореневого вузла S, який містить повний набір даних.

2. Знайти найкращий атрибут у наборі даних шляхом вимірювання вибору атрибутів.

3. Розділити S на підмножини, що містять можливі значення для найкращих атрибутів.

4. Створити вузол дерева рішень, який містить найкращий атрибут.

5. Рекурсивно створити нові дерева рішень, використовуючи підмножини набору даних, створеного на кроці 3. Продовжувати цей процес досягнення стадії, коли вузли не можна далі класифікувати і вони стають листовими вузлами [49].

Під час реалізації дерева рішень постає основне питання щодо вибору найкращого атрибута для кореневого вузла та підвузлів. Тому існує методика, відома як міра вибору атрибутів, яка допомагає вирішити такі проблеми. За допомогою цього вимірювання можна легко вибрати найкращий атрибут для вузлів дерева. Для досягнення цієї мети використовуються популярні методи:

1. Коефіцієнт приросту інформації: використовується для класифікації. Алгоритм дерева рішень завжди намагається максимізувати значення цього коефіцієнта, тому спочатку вибирається вузол/атрибут з найбільшим інформаційним коефіцієнтом.

2. Індекс Джині: використовується для класифікації. Варто віддавати перевагу атрибуту з найнижчим значенням індексу Джині порівняно з високим значенням індексу Джині.

3. Метод найменших квадратів: використовується для регресії. Щоб визначити "найкращий" розподіл, потрібно обчислити суму квадратів залишків (RSS) для кожного предиктора, порівняти їх значення та знайти найнижче.

Математично RSS (сума квадратів залишків) можна записати таким чином за формулою (2.23) [50]:

$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (2.23)$$

Ансамблеве навчання є процесом, у якому кілька класифікаторів комбінуються для розв'язання складної задачі та підвищення ефективності моделі.

Випадковий ліс є одним з методів ансамблевого навчання. Цей класифікатор складається з набору дерев рішень, які будуються на підмножинах даних, та використовує середнє значення прогнозів цих дерев для покращення точності прогнозування. Замість використання лише одного дерева рішень, випадковий ліс отримує прогнози з кожного дерева і, шляхом усереднення цих прогнозів, дає кінцевий результат [51].

Випадковий ліс пов'язаний з поняттями бутстрапінгу та беггінгу. Бутстрапінг є методом вибірки, де кожне спостереження вибирається з набору заміщенням, що означає, що деякі спостереження можуть бути використані кілька разів у одному дереві. Потім алгоритм навчання застосовується до вибраних вибірок. Ця техніка використовує вибірку заміщенням, щоб зробити процедуру вибору абсолютно випадковою. Беггінг складається з двох складових: агрегації та бутстрапінгу. Цей метод об'єднує кілька слабких моделей, агрегуючи їх прогнози для вибору найкращого [52].

Процес створення випадкового лісу складається з таких етапів:

1. Випадковим чином виберіть k об'єктів спостережень з загальної кількості z, від b = 1 до B.

2. Зростіть дерево випадкового лісу на основі вибраних даних. Для кожного термінального вузла дерева повторюйте наступні кроки, доки не досягнутий мінімальний розмір вузла:

   a. Випадковим чином виберіть m змінних з p змінних.

   b. Виберіть найкращу змінну / точку розбиття серед m.

c. Розділіть вузол на дочірні вузли.

d. Повторіть кроки I-III для новостворених вузлів, якщо це необхідно.

3. Отриманий ансамбль дерев'їв позначається як$\{T_b\}_1^B$.

4. Для здійснення прогнозу для регресії у новій точці x використовуйте формулу (2.24) [40, с. 603].

$$\hat{f}_{rf}^B(x) = \frac{1}{B} * \sum_{b=1}^{B} T_b(x), \quad (2.24)$$

де B – кількість дерев.

Важливо усвідомлювати, що випадковий ліс не може проводити екстраполяцію. Це означає, що коли перед випадковим лісовим регресором ставиться завдання прогнозування значень, які раніше не спостерігалися, він завжди буде прогнозувати середнє значення з раніше спостережених значень. Очевидно, що середнє значення вибірки не може виходити за межі найвищих і найнижчих значень у вибірці [53]. Щоб оцінити, наскільки добре працює випадковий ліс у завданнях класифікації, потрібно використовувати оцінку помилки Out-of-Bag (OOB). При побудові кожного дерева рішень у лісі, ми використовуємо підвибірку записів (тобто бутстрепінг), і залишок записів можна використовувати для тестування цього дерева і обчислення рівня помилок. Таким чином, коефіцієнт помилок OOB є синтетичним показником точності моделі [54]. Щоб отримати оцінку OOB, потрібно:

1. Знайти всі записи, які не брали участь у побудові дерев рішень (їх називають "OOB дані").

2. Використовувати ці записи для тестування на всіх деревах рішень у випадковому лісі.

3. Зрештою, можна виміряти, наскільки точним є випадковий ліс, шляхом визначення частки OOB записів, які були правильно класифіковані випадковим лісом; ця частка помилково класифікованих записів називається "похибка OOB" [61].

2.4 Опис методології побудови алгоритму XGBoost

У розділі 2.2 було розглянуто суть ансамблевого навчання, де два найпопулярніших методи - беггінг та бустинг. Беггінг використовується в алгоритмі випадкового лісу, де кілька окремих моделей навчаються паралельно на випадкових підмножинах даних. З іншого боку, бустинг будує багато окремих моделей послідовно, кожна з яких навчається на помилках, допущених попередньою моделлю [56]. Техніка бустингу включає три прості кроки:

1. Початкова модель F0 використовується для прогнозування цільової змінної У. Ця модель асоційована з залишком (у - F0).

2. Нова модель h1 побудована для передбачення залишків від попереднього кроку.

3. Зараз F0 і h1 комбінуються, щоб отримати F1, покращену версію F0. Середня квадратична помилка з F1 буде нижчою, ніж з F0, що випливає з рівняння (2.25):

$$F_1(x) \prec F_0(x) + h_1 \tag{2.25}$$

Можна було б покращити продуктивність F1, створивши нову модель F2, що враховує залишки F1, згідно з рівнянням (2.26):

$$F_2(x) \prec F_1(x) + h_2(x) \tag{2.26}$$

Можна провести "m" ітерацій до того, як залишки будуть зменшені настільки, наскільки це можливо, як показано у рівнянні (2.27) [57].

$$F_m(x) \prec F_{m-1}(x) + h_m(x) \tag{2.27}$$

Кінцевою моделлю тут є поетапна адитивна модель b окремих дерев, як видно у формулі (2.28):

$$f(x) = \sum_{b=1}^{B} f^b(x) \tag{2.28}$$

Градієнтний бустинг, який є одним з видів бустингу в машинному навчанні, базується на ідеї градієнтного спуску. Основна мета градієнтного спуску - ітеративно налаштовувати параметри з метою мінімізації функції витрат.

Градієнтний бустинг використовує логіку, що краща наступна модель, яка поєднується зі зведеними моделями, допомагає мінімізувати загальну помилку прогнозування.

У випадку задач регресії ми використовуємо середньоквадратичну помилку (MSE) як метрику оцінки, а для задач класифікації - логарифмічні втрати [58].

Основний принцип градієнтного бустингу полягає в мінімізації цільової функції за допомогою такого підходу: на кожній ітерації ми шукаємо «базову модель» шляхом підбору від'ємного градієнта функції втрат, потім множимо цей прогноз на константу і додаємо до значення попередньої ітерації.

Умови задачі: навчальний набір $\{(x_i, y_i)\}_{i=1}^{n}$, функція втрат L(y,F(x)) яка є диференційованою, кількість ітерацій M.

Алгоритм реалізації моделі за допомогою градієнтного бустингу:

1. Ініціюємо модель з постійним значенням (2.29):

$$F_0(x) = argmin_\gamma \sum_i^n L(y_i, \gamma) \tag{2.29}$$

2. З m = 1 до M:

a. Градієнт функції втрат (псевдо-залишки) обчислюється ітеративно (2.30):

$$r_{im} = -\alpha \left[\frac{\partial L(y_i, F(x_i)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \tag{2.30}$$

де α – швидкість навчання.

b. Кожна «базова модель» (наприклад дерево) $h\,m(x)$ відповідає градієнту, отриманому на кожному кроці $\{(x_i, r_{im})\}_{i=1}^{n}$;

c. множник γm вирішивши наступну одновимірну задачу оптимізації за допомогою формули (2.31):

$$\gamma_m = argmin_\gamma \sum_i^n L(y_i, F_{m-1}(x) + \gamma h_m(x_i)) \tag{2.31}$$

d. Оновити модель за допомогою формули (2.32):

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2.32}$$

е. Вивести $F_M(x)$.

Інтуїція полягає у тому, що за допомогою градієнтного бустингу ми здійснюємо градієнтний спуск на функцію втрат, підбираючи "base learner" до негативного градієнта на кожній ітерації. Негативні градієнти, які також називають псевдо-залишками, допомагають мінімізувати цільову функцію [59]. Градієнтний бустинг навчається безпосередньо на залишковій помилці, а не оновлює ваги точок даних.

Більш інтуїтивний алгоритм навчання для регресії:

1. Зробити початкове припущення, що полягає в прогнозуванні середнього значення цільового Y. Зазвичай, початкове припущення розраховується як середнє значення Y.

2. Обчислити псевдо-залишки, віднімаючи початкове припущення від спостережень. Це дає нам псевдо-залишки.

3. Побудувати дерево рішень, використовуючи псевдо-залишки як цільову змінну.

4. Зробити прогноз і обчислити нові залишки. Градієнтний бустинг використовує коефіцієнт навчання α для масштабування внеску нового дерева, застосовуючи коефіцієнт від 0 до 1.

5. Повторити наступні кроки:

I. Побудувати наступне дерево, використовуючи залишки, обчислені на кроці 4.

II. Зробити прогноз, використовуючи всі побудовані дерева.

III. Обчислити нові залишки відповідно до прогнозу.

IV. Продовжувати будувати дерева до досягнення певного критерію зупинки [54].

XGBoost, також відомий як екстремальний градієнтний бустинг, є одним з відомих методів градієнтного бустингу, який покращує продуктивність і

швидкість алгоритмів машинного навчання, заснованих на деревах (послідовних деревах рішень) [60]. Особливості XGBoost включають:

1. Регуляризація: XGBoost накладає штраф на складні моделі, використовуючи як L1 (лассо), так і L2 (рідж) регуляризацію, щоб запобігти перенавчанню.

2. Перехресна перевірка: У XGBoost вбудований метод перехресної перевірки, який автоматично використовується на кожній ітерації. Це усуває потребу в ручному налаштуванні кількості посилення ітерацій для одного запуску моделі.

3. Гнучкість: XGBoost підтримує різні цільові функції і може працювати з визначеними користувачем метриками оцінки.

4. Обрізка дерева: Це техніка машинного навчання, яка дозволяє скоротити розмір регресійних дерев шляхом видалення вузлів, які не сприяють поліпшенню класифікації на листі. Мета обрізки дерева полягає в запобіганні перенавчанню під час навчання моделі [61].

Алгоритм XGBoost для регресії працює наступним способом:

1. Першим кроком у пристосуванні XGBoost до навчальних даних є виконання початкового прогнозу. За замовчуванням це значення рівне 0,5, але може мати будь-яке інше значення.

2. Залишки, що представляють різницю між спостережуваними значеннями і прогнозами, вказують, наскільки точним є початковий прогноз. Ми можемо кількісно оцінити якість прогнозу за допомогою функції втрат, формула (2.33) якої наведена нижче. Пізніше, ми можемо застосувати цю функцію втрат до нових прогнозів і порівняти результати, щоб визначити, чи поліпшуються прогнози.

$$L(y_i, p_i) = \frac{1}{2}(y_1 - p_1)^2 + \cdots + \frac{1}{2}(y_n - p_n)^2, \quad (2.33)$$

де y_i – спостережені значення;

p_i – прогнозовані значення, які відповідають кожному із спостережень y_i.

3. Після цього, XGBoost використовує функцію втрат для побудови дерев, мінімізуючи рівняння (2.34):

$$[\sum_{i=1}^{n} L(y_i, p_i)] + \gamma T + \frac{1}{2}\lambda w^2 \quad (2.34)$$

де $[\sum_{i=1}^{n} L(y_i, p_i)]$ – функція втрат;

γ – штраф, який визначається користувачем, щоб заохочення обрізки дерева;

T – кількість термінальних вузлів, або листків у дереві;

$\frac{1}{2}\lambda w^2$ – термін регуляризації.

Мета полягає в пошуку вихідного значення (w) для листка, яке мінімізує всі рівняння. Оскільки ми плануємо оптимізувати вихідне значення з першого дерева, можна замінити прогнозоване значення (p_i) на початкове прогнозоване значення p^0 і додати вихідне значення (w) нового дерева. Це можна виразити у формулі (2.35):

$$[\sum_{i=1}^{n} L(y_i, p^0 + w)] + \gamma T + \frac{1}{2}\lambda w^2 \quad (2.35)$$

4. Далі необхідно знайти вихідне значення (w) яке мінімізує рівняння (2.33). Чим більше ми застосовуємо регуляризацію, збільшуючи значення λ, тим більше вихідне значення (w) наближається до нуля. Вихідне значення (w) можна знайти за формулою (2.36):

$$w = \frac{Sum\ of\ Residuals}{(Number\ of\ Residuals + \lambda)} \quad (2.36)$$

З практичного погляду, опис алгоритму XGBoost для регресії можна узагальнити наступним чином:

1. Перший етап полягає у групуванні всіх залишків в одному термінальному вузлі (листку).

2. Далі ми використовуємо рівняння (2.37) для визначення показника подібності (similarity score), що дозволяє нам будувати дерево.

$$Similarity\ score = \frac{RSS}{(Number\ of\ Residuals + \lambda)}, \quad (2.37)$$

де λ – параметр регуляризації.

3. Значення порогу для кореня дерева обчислюється як середнє значення двох близьких точок при розбитті, а залишок йде у відповідний листок.

4. Оцінка схожості розраховується для обох вузлів дерева. Якщо залишки в вузлі значно відрізняються, вони компенсують один одного, що призводить до дуже малої оцінки схожості, і навпаки.

5. Далі визначаємо, наскільки краще вузли дерева кластеризують залишки порівняно з коренем дерева. Це робиться шляхом розрахунку коефіцієнта посилення (Gain) за формулою (2.38):

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\alpha} + \frac{G_R^2}{H_R+\alpha} - \frac{(G_L+G_R)^2}{H_L+H_R+\alpha}\right] - \gamma \qquad (2.38)$$

6. Ми будуємо інші варіанти дерева аналогічним чином, розділяючи спостереження за допомогою нових порогових значень. Для кожного варіанту розраховуємо оцінку схожості та коефіцієнт посилення. Порівнюємо значення коефіцієнта посилення для дерев з різними пороговими значеннями. Перше гілку дерева буде визначено за допомогою порогового значення, яке має найбільше значення коефіцієнта посилення.

7. Далі, за тим самим принципом, можна розділяти залишки в кожному термінальному вузлі. Глибину дерева може визначати користувач, але за замовчуванням ми дозволяємо до 6 дерев.

8. Після побудови дерев XGBoost можна виконати необов'язковий крок обрізки дерева [див. формулу (2.39)]. Обрізка полягає у видаленні обраних гілок з дерева. Мета полягає в усуненні небажаних гілок, поліпшенні структури дерева та спрямуванні його на нове здорове зростання. Обрізка дерева здійснюється на основі значення коефіцієнта посилення та параметра γ, який задається користувачем. Цей процес розпочинається знизу (на рівні листя) і продовжується до кореневого вузла, перевіряючи, чи не стає значення коефіцієнта посилення менше заданого значення γ.

$$\text{Gain} - \gamma \tag{2.39}$$

У разі, коли від'ємна різниця між коефіцієнтом підсилення та параметром γ спостерігається, гілка дерева буде видалена, і навпаки, у випадку протилежного значення різниці.

9. Для завершення побудови дерева необхідно обчислити вихідне значення (w) за допомогою формули (2.36) для кожного листка дерева.

10. Після побудови першого дерева можна здійснити новий прогноз. Подібно до звичайного градієнтного бустингу, XGBoost здійснює нові прогнози, стартуючи зі початкового передбачення і додаючи вихідне значення (w), помножене на коефіцієнт навчання ε (за замовчуванням 0.3).

11. Нові дерева будуються на основі прогнозованих залишків і дають прогнози з ще меншими залишками. Ми продовжуємо будувати дерева, поки залишки не стануть дуже малими або досягнемо максимальної кількості дерев [62].

РОЗДІЛ 3. ПРАКТИЧНЕ ЗАСТОСУВАННЯ ОПИСАНИХ МЕТОДІВ ДЛЯ ПОБУДОВИ ПРОГНОЗУ

3.1 Розвідувальний аналіз даних

Прогнозування вартості нерухомості виступає ключовим економічним фактором для учасників ринку нерухомості, у тому числі для:

1. Майбутніх покупців або власників нерухомості: Люди, які планують купити або інвестувати в нерухомість, цікавляться прогнозом цін, щоб зробити осмислені рішення про покупку та оптимальний час для цього.

2. Забудовники та розробники: Компанії, які займаються будівництвом та розвитком нерухомості, потребують точного прогнозу цін для планування нових проектів, визначення прибутковості та стратегій розширення.

3. Інвестори: Інвестори в нерухомість вкладають гроші з метою отримання прибутку. Точний прогноз ціни допомагає їм зробити розумний вибір щодо вкладення коштів.

4. Оцінювачі нерухомості: Оцінювачі визначають ринкову вартість нерухомості. Для цього вони використовують прогнози цін на нерухомість, які допомагають їм зробити об'єктивну оцінку.

5. Податкові департаменти: Органи оподаткування використовують прогнози цін на нерухомість для встановлення податкових ставок, обчислення податків на власність та планування бюджету.

6. Іпотечні кредитори та страховики: Банки, кредитні установи та страхові компанії враховують прогнози цін на нерухомість при видачі іпотечних кредитів та встановленні страхових тарифів. Наявність моделі прогнозування ціни на житло допомагає заповнити прогалину в інформації та підвищує ефективність ринку нерухомості.

Ціна на нерухомість залежить від різних факторів, які були вже відображені в першому розділі роботи. Для прогнозування цін на нерухомість в окрузі Кінг у США в цьому розділі буде використано набір даних, що містить

різноманітні характеристики цього регіону, з метою створення моделей. Джерелом інформації для дослідження є дані з ресурсу https://www.kaggle.com/harlfoxem/housesalesprediction [63].

Набір даних для аналізу складається з цін на будинки, проданих протягом періоду з травня 2018 року по травень 2019 року. Разом з ціною будинку, набір даних містить інформацію про характеристики будинку, які можна знайти у таблиці 3.1.

Таблиця 3.1 – Набір даних та значення полів [64]

Показник	Опис показника
Id	Ідентифікатор будинку
Date	Дата продажу
Price	Ціна продажу
Bedrooms	Кількість спальних кімнат
Bathrooms	Кількість ванних кімнат
Sqft_living	Площа житлового простору будинків у квадратних футах
Sqft_lot	Площа земельного простору будинків у квадратних футах (площа ділянки)
Floors	Кількість поверхів
Waterfront	Фіктивна змінна: чи має будинок вид на набережну
View	Індекс від 0 до 4, що вказує на якість виду з будинку
Condition	Індекс від 1 до 5, що оцінює стан будинку
Grade	Індекс від 1 до 13, що відображає рівень будівельної та дизайнерської якості
Sqft_above	Площа внутрішнього простору житла, що знаходиться над рівнем землі, у квадратних футах
Sqft_basement	Площа внутрішнього простору, що знаходиться під рівнем землі, у квадратних футах
Yr_built	Рік побудови
Yr_renovated	Рік ремонту
Zipcode	Поштовий індекс району, в якому знаходиться будинок
Lat	Широта
Long	Довгота

Оскільки в наборі даних присутні змінні, що мають значення індексів, потрібно розглянути кожну з них окремо для їх інтерпретації.

Таблиця 3.2 містить значення кожного індексу показника "view" (як добре видно з будинку).

Таблиця 3.2. Значення індексів, які характеризують якість виду з будинку [65].

Показник View	Значення
0	Невідомо
1	Задовільний
2	Прийнятний
3	Гарний
4	Відмінний

У табл. 3.3 наведено значення кожного індексу показника condition (стан будинку).

Таблиця 3.3 – Значення індексів показника condition (стан будинку) [65]

Показник Condition	Значення	Опис
1	Погане	Останнім часом стан будинку серйозно погіршився і потребує багатьох ремонтів.
2	Задовільне	Негайно потрібні деякі ремонтні роботи, а також відкладене обслуговування.
3	Прийнятне	В залежності від часу проведення ремонтних робіт, це вважається нормальним станом для будинку такого віку.
4	Чудове	Стан будинку перевищує норму для його віку, що свідчить про додаткову увагу та обережність у підтримці.
5	Відмінне	Будинок відмінно обслуговується та оновлюється.

У таблиці А (див. Додаток А) містяться значення кожного індексу показника "grade", що відображає відповідність будівельному та дизайнерському рівням [66].

Спочатку, необхідно провести аналіз початкових даних, що є першим кроком для будь-якого дослідження. Для цього були використані програмні засоби, такі як мова програмування R та програма Power BI. Після завантаження даних у вказані програми, можна зробити певні висновки. На рисунку 3.1 наведено перші 10 спостережень для всіх відображених стовпців.

	id	price	date	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long
1	1	221900	2018-10-13	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257
2	2	538000	2018-12-09	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951	1991	98125	47.7210	-122.319
3	3	180000	2019-02-25	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233
4	4	604000	2018-12-09	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393
5	5	510000	2019-02-18	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045
6	6	1225000	2018-05-12	4	4.50	5420	101930	1.0	0	0	3	11	3890	1530	2001	0	98053	47.6561	-122.005
7	7	257500	2018-06-27	3	2.25	1715	6819	2.0	0	0	3	7	1715	0	1995	0	98003	47.3097	-122.327
8	8	291850	2019-01-15	3	1.50	1060	9711	1.0	0	0	3	7	1060	0	1963	0	98198	47.4095	-122.315
9	9	229500	2019-04-15	3	1.00	1780	7470	1.0	0	0	3	7	1050	730	1960	0	98146	47.5123	-122.337
10	10	323000	2019-03-12	3	2.50	1890	6560	2.0	0	0	3	7	1890	0	2003	0	98038	47.3684	-122.031

Рисунок 3.1 – Перші десять спостережень [63]

У мові програмування R можна провести пошук значень NaN, Na та NULL. NA є логічною константою довжиною 1, яка вказує на відсутнє значення. NULL використовується для представлення пустого об'єкта в R. NaN позначає "Not a Number" (не число). За результатами, які показані на рисунку 3.2, можна стверджувати, що не було знайдено значень з такими характеристиками. Це важливий сигнал для аналізу даних.

```
> is.null(data)
[1] FALSE
> sum(is.na (data))
[1] 0
> sum(is.nan(as.matrix(data)))
[1] 0
```

Рисунок 3.2 – Пошук значень NaN, Na та NULL

За допомогою функції dim() у мові програмування R було отримано розмірність даних, яку можна побачити на рис. 3.3. З отриманого результату можна зробити висновок, що дані мають 19 показників та 21613 спостережень.

```
> #Потрібно перевірити розмірність датафрейму
> dim(data)
[1] 21613    19
```

Рисунок 3.3 – Розмірність даних

За допомогою функції str() у мові програмування R було використано для визначення структури даних, зображених на рис. 3.4. З цього результату можна зробити висновок, що всі змінні, крім дати, мають числовий тип даних, який включає цілі числа і дробові числа. Дата, натомість, належить до класу дати та часу - POSIXct.

```
tibble [21,613 x 19] (S3: tbl_df/tbl/data.frame)
 $ id           : num [1:21613] 1 2 3 4 5 6 7 8 9 10 ...
 $ price        : num [1:21613] 221900 538000 180000 604000 510000 ...
 $ date         : POSIXct[1:21613], format: "2018-10-13" "2018-12-09" "2019-02-25" "2018-12-09" ...
 $ bedrooms     : num [1:21613] 3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms    : num [1:21613] 1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living  : num [1:21613] 1180 2570 770 1960 1680 ...
 $ sqft_lot     : num [1:21613] 5650 7242 10000 5000 8080 ...
 $ floors       : num [1:21613] 1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront   : num [1:21613] 0 0 0 0 0 0 0 0 0 0 ...
 $ view         : num [1:21613] 0 0 0 0 0 0 0 0 0 0 ...
 $ condition    : num [1:21613] 3 3 3 5 3 3 3 3 3 3 ...
 $ grade        : num [1:21613] 7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above   : num [1:21613] 1180 2170 770 1050 1680 ...
 $ sqft_basement: num [1:21613] 0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built     : num [1:21613] 1955 1951 1933 1965 1987 ...
 $ yr_renovated : num [1:21613] 0 1991 0 0 0 ...
 $ zipcode      : num [1:21613] 98178 98125 98028 98136 98074 ...
 $ lat          : num [1:21613] 47.5 47.7 47.7 47.5 47.6 ...
 $ long         : num [1:21613] -122 -122 -122 -122 -122 ...
```

Рисунок 3.4 – Структура даних

Оскільки в наборі змінних присутній рік будівництва та рік останнього ремонту, було вирішено додати дві додаткові змінні, які відображають вік будинку та вік останнього ремонту. Ці показники допоможуть краще зрозуміти дані та здійснити моделювання. Важливим кроком є проведення кореляційного аналізу для дослідження залежностей між змінними. Для наглядності, можна побудувати матрицю кореляцій, яку можна побачити на рисунку 3.5.

Аналізуючи цей графік, ми можемо зробити висновок, що деякі незалежні змінні сильно корелюють між собою, що може свідчити про наявність мультиколінеарності. Мультиколінеарність виникає, коли незалежна змінна має сильну кореляцію з однією або кількома іншими незалежними змінними в рівнянні множинної регресії. Ця проблема є важливою, оскільки вона підриває статистичну значимість незалежної змінної.

Також, можна зробити висновок, що найсильніший зв'язок з ціною (кореляція більше 0.5) спостерігається у таких показниках: площа житла

(sqft_living), класифікація (grade), площа надземної частини (sqft_above) та кількість ванних кімнат (bathrooms).

Відношення між ціною та площею житла складає 0.7, що вказує на тісний прямий зв'язок - чим більша площа будинку, тим вища ціна. Відношення між ціною та класифікацією становить 0.67, що свідчить про середній прямий зв'язок - чим вищий рейтинг, тим вища ціна. Відношення між ціною та площею надземної частини дорівнює 0.61, що вказує на середній прямий зв'язок - чим більша площа, тим вища ціна. Відношення між ціною та кількістю ванних кімнат складає 0.53, що свідчить про середній прямий зв'язок - чим більше ванних кімнат, тим вища ціна. З іншого боку, найнижчий рівень кореляції з ціною спостерігається у показнику "дата", який дорівнює нулю.

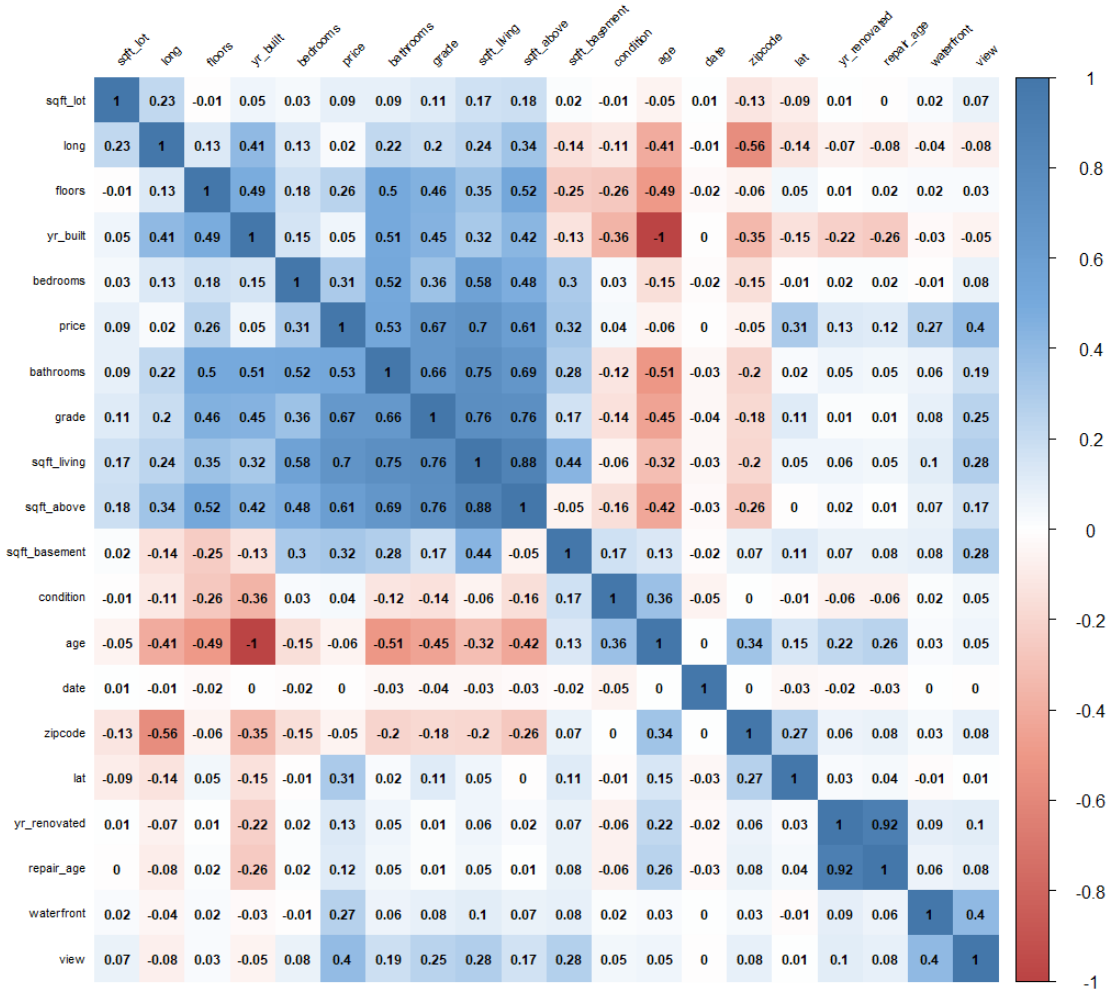


Рисунок 3.5 – Кореляційна матриця

Після завантаження даних до програми Power BI, можна отримати висновки шляхом візуалізації. З метою більш детального аналізу, було вирішено встановити відповідність між поштовими кодами та містами, які вони охоплюють. Таким чином, на рисунку 3.6 представлені 70 унікальних поштових кодів та відповідні міста. Слід зауважити, що в окремих випадках один поштовий код може охоплювати кілька міст.

zipcode	Місто	zipcode	Місто
98001	Auburn	98070	Vashon
98002	Auburn	98072	Woodinville
98003	Federal Way	98074	Sammamish
98004	Bellevue	98075	Sammamish
98005	Bellevue	98077	Woodinville
98006	Bellevue	98092	Auburn
98007	Bellevue	98102	Seattle
98008	Bellevue	98103	Seattle
98010	Black Diamond	98105	Seattle
98011	Bothell	98106	Seattle
98014	Carnation	98107	Seattle
98019	Duvall	98108	Seattle
98022	Enumclaw	98109	Seattle
98023	Federal Way	98112	Seattle
98024	Fall City	98115	Seattle
98027	Issaquah	98116	Seattle
98028	Kenmore	98117	Seattle
98029	Issaquah	98118	Seattle
98030	Kent	98119	Seattle
98031	Kent	98122	Seattle
98032	Kent	98125	Seattle
98033	Kirkland	98126	Seattle
98034	Kirkland	98133	Seattle
98038	Maple Valley	98136	Seattle
98039	Medina	98144	Seattle
98040	Mercer Island	98146	Seattle
98042	Kent	98148	Seattle
98045	North Bend	98155	Seattle
98052	Redmond	98166	Seattle
98053	Redmond	98168	Seattle
98055	Renton	98177	Seattle
98056	Renton	98178	Seattle
98058	Renton	98188	Seattle
98059	Renton	98198	Seattle
98065	Snoqualmie	98199	Seattle

Рисунок 3.6 – Відповідність поштових кодів до міст

На рисунку 3.7 представлені три гістограми, що відображають розподіл будинків за такими характеристиками: кількість поверхів, кількість спалень і кількість ванних кімнат. Можливо, виникає питання, чому кількість ванних кімнат виражена десятковими числами. Пояснення полягає в тому, що в Сполучених Штатах ванні кімнати зазвичай класифікуються як: основна ванна кімната (містить душ і ванну), ванна кімната, що прилягає до спальні, "повна"

ванна кімната (з чотирма сантехнічними пристроями: ванна, душ, туалет і раковина), "половина" ванни (з туалетом і раковиною) та "3/4" ванни (з унітазом, раковиною і душем), хоча ці визначення можуть відрізнятися в залежності від ринку [67].

Також важливо відзначити, що в США перший поверх означає поверх, що знаходиться на рівні землі, другий поверх - поверх над ним і так далі. Горище, підвал і подібні приміщення вважаються нецілими поверхами.

Аналізуючи рисунок, можна зробити висновок, що найбільша кількість будинків (10680) має один поверх, також є багато будинків з двома поверхами (8241), а кількість будинків з 3,5 поверхами становить найменше (8). Щодо кількості спалень, найбільше будинків (9824) мають 3 спальні, також часто зустрічаються будинки з 4 спальнями (6882), а найменша кількість будинків мають 10, 11 або 33 спальні.

Розподіл за ванними кімнатами вказує на те, що найбільшу кількість будинків (10285) мають від 1 до 2 ванних кімнат, а також від 2 до 3 ванних кімнат (9365). За своєю чергою, найменшу кількість будинків (4) спостерігається в діапазоні від 7 до 8 ванних кімнат.

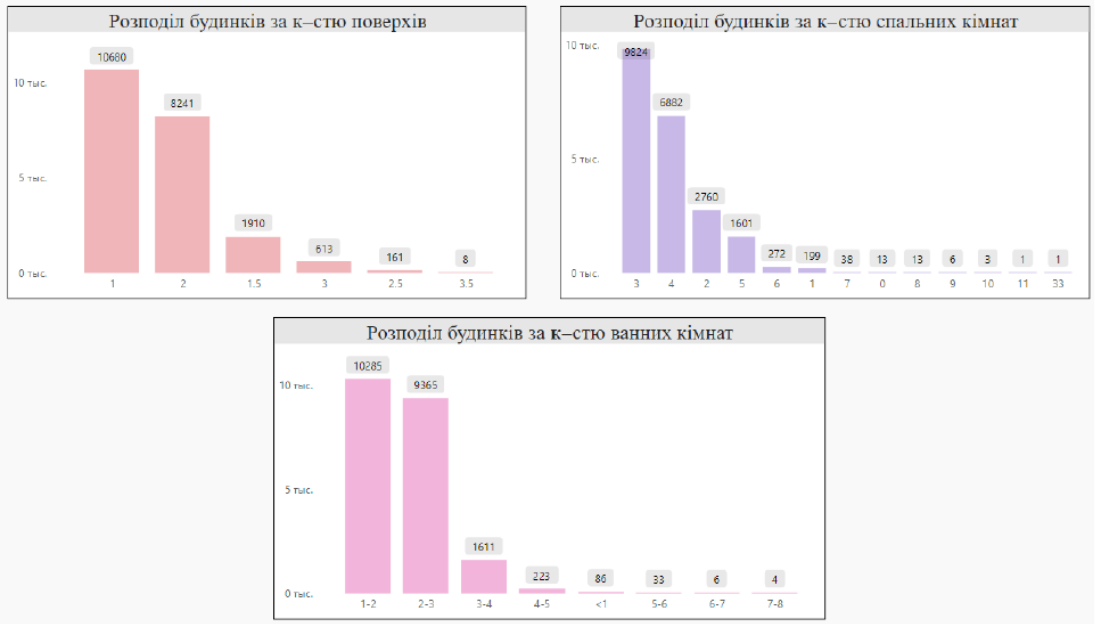


Рисунок 3.7 – Розподіл будинків за кількістю кімнат

На рисунку 3.8 представлено чотири гістограми, які показують розподіл будинків за наступними характеристиками: стан будинку, вид будинку, відповідність будівельному та дизайнерському рівням, а також розташування нанабережну. За даними графіками можна зробити такі висновки: найбільша кількість будинків (14031) має індекс 3, що вказує на нормальний стан для будинку такого віку, а найменша кількість будинків (30) має індекс 1, що свідчить про дуже малий відсоток будинків, яким потрібні значні ремонтні роботи.

Щодо індексу виду будинку, найбільша кількість будинків (19489) має індекс 0, що означає відсутність інформації, а найменша кількість будинків (319) має індекс 4. Розподіл за індексом відповідності будівельному та дизайнерському рівням показує, що найбільша кількість будинків (8981) має індекс 7, що вказує на середню оцінку конструкції і дизайну, а найменша кількість будинків має індекси 1 (1) та 3 (3), що свідчить про те, що ці будинки не відповідають мінімальним будівельним стандартам. Також можна зробити висновок, що жоден будинок не має індексу 2.

За даними про розміщення будинків відносно набережної, видно, що більшість будинків (21450) не мають виходу прямо на набережну, в той час як лише 163 будинки безпосередньо примикають до неї.

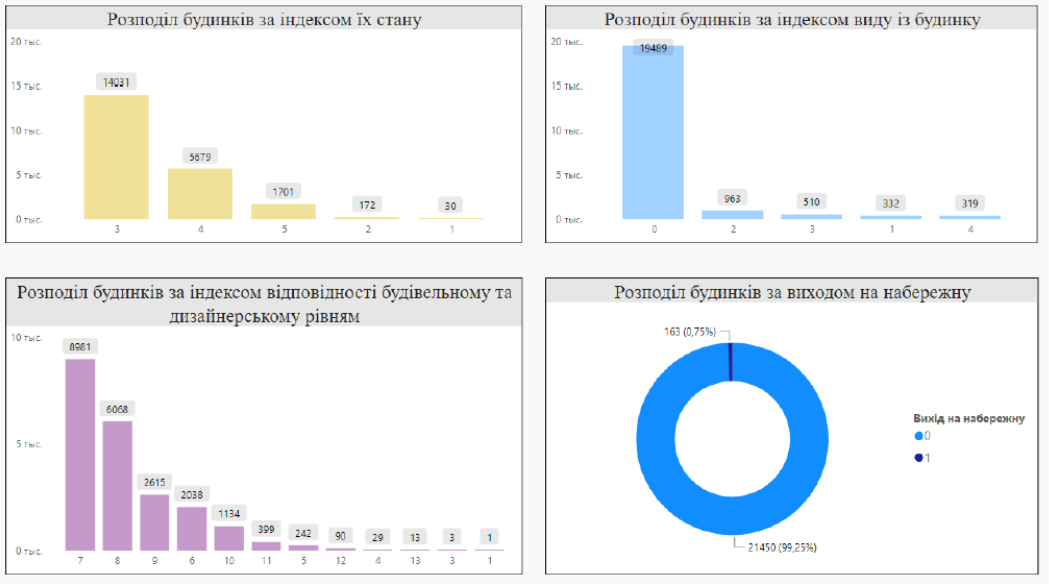


Рисунок 3.8 – Розподіл будинків за індексами

На рисунку 3.9 представлена інформація про розташування будинків. Зазначено, що найбільша кількість будинків знаходиться у місті Сіетл (8977). Крім того, значна кількість будинків знаходиться в містах Рентон, Беллав'ю та Кент (відповідно 1597, 1407 та 1203 будинків). Найменша кількість будинків є в містах Фолл Сіті та Медіна. Ці розбіжності можна пояснити тим, що, наприклад, у місті Сіетл проживає 724 305 осіб, тоді як у Фолл Сіті та Медіні нараховується менше 4 тисяч осіб.

Рисунок 3.9 – Розподіл будинків за містами їх знаходження

На рисунку 3.10 показано, як розподіляються будинки в залежності від їх житлової площі та площі ділянки. Житлова площа та площа ділянки були розділені на приблизно однакові групи, щоб полегшити їх розуміння та візуалізацію. Зробивши аналіз, можна зробити висновок, що більшість будинків мають площу в діапазоні від 1500 до 2000 (5382), від 1000 до 1500 (4836) та від 2000 до 2500 (4204). Натомість, найдекілька будинків мають найбільші площі, які знаходяться в діапазоні від 6500 до 7000 (15), від 7000 до 7500 (13), понад

8000 (9) та від 7500 до 8000 (6). Іншими словами, більшість будинків мають невелику площу.

Можна зробити висновок, що більшість будинків мають площу менше 10000 (15443, а також площу в діапазоні від 10000 до 20000 (3858). Значна кількість будинків мають середню та велику площу.

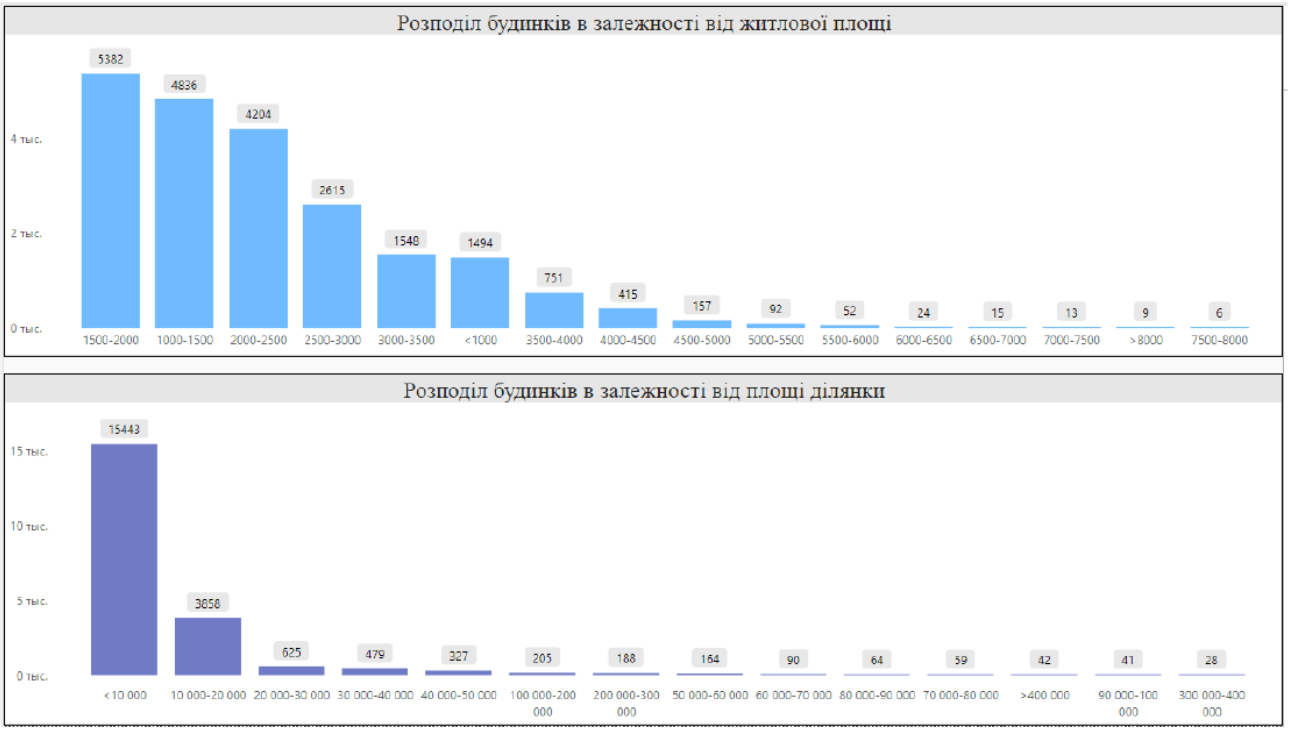


Рисунок 3.10 – Розподіл будинків в залежності від площі

На рисунку 3.11 представлений графік, що демонструє розподіл будинків в залежності від їх віку та віку ремонту. Для кращого сприйняття і візуалізації, вік був розбитий на декілька рівних груп. Аналізуючи результати, ми можемо зробити наступні висновки: найбільша кількість будинків мають вік від 11 до 20 років (3443), приблизно однакова кількість будинків знаходиться віком від 21 до 60 років.

Мінімальна кількість будинків належить до дуже старих, вік яких коливається від 81 до 90 років (596), а також до дуже молодих будинків, побудованих менше 10 років тому (1098). Найбільша кількість будинків має вік ремонту понад 100 років (229), мінімальна кількість належить до ремонту,

проведеного віком від 21 до 30 років (5), і тільки один будинок має вік ремонту від 11 до 20 років. Крім того, не було знайдено жодного будинку, в якому ремонт був проведений останніми 10 роками.

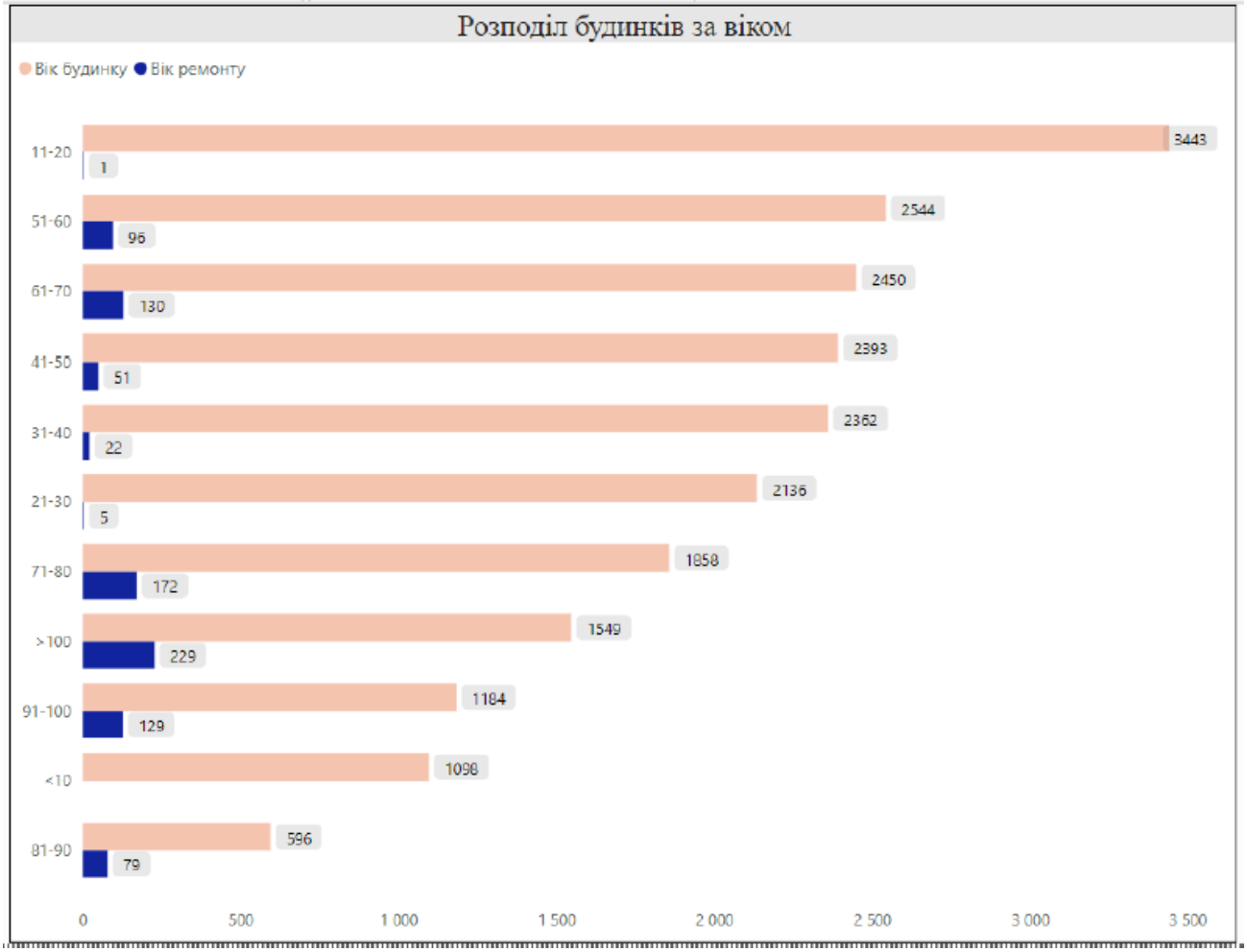


Рисунок 3.11 – Розподіл будинків за віком

Розподіл будинків в залежності від років будівництва і ремонту показано на рис. 3.12. Роки були розділені на групи для кращого розуміння і візуалізації. Зробивши аналіз, можна зробити висновок, що найбільша кількість будинків була побудована у період з 2001 по 2010 рік (3443), в той час як кількість будинків, побудованих до 1920 року, досить невелика. Найменша кількість будинків була побудована у період з 1931 по 1940 рік, що можна пояснити Великою Депресією, яка тривала у США з 1929 по 1939.

Більшість будинків була піддана ремонту у період з 1940 по 1960 роки, а також у 1921–1930 роках (129). Найменша кількість будинків мала ремонт, зроблений у період з 1990 до 2010.

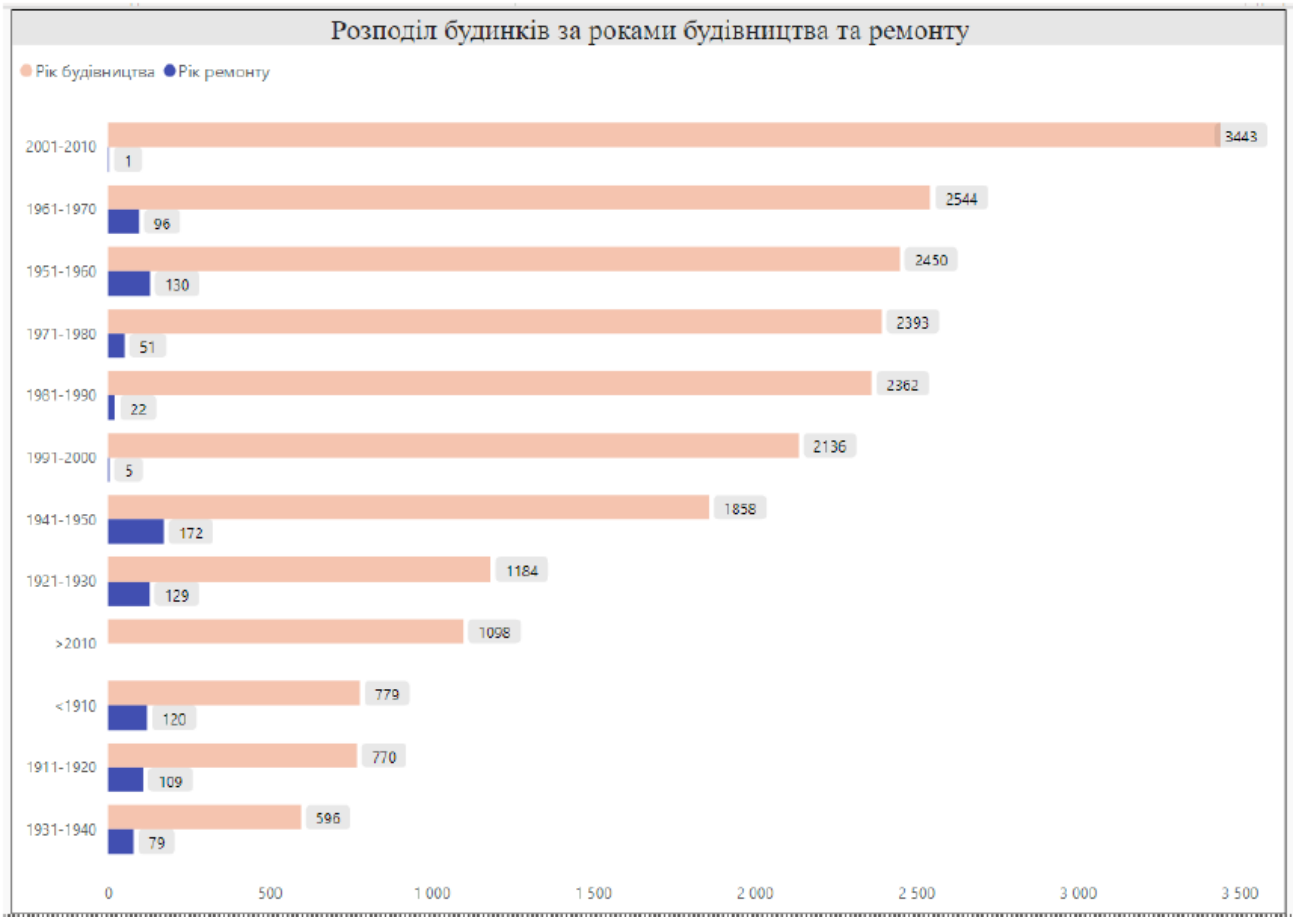


Рисунок 3.12 – Розподіл будинків за роками будівництва та ремонту

На рисунку 3.13 показано, як залежить кількість проданих будинків від дати продажу. Зробивши аналіз, можна зробити висновок, що найбільше число будинків, а саме 2414, було продано у травні в період від 2018 до 2019 року, а найменше число, а саме 978, було продано у січні. Можна відзначити тенденцію, що більше будинків було продано в теплий період року, а з настанням холодів кількість продажів поступово зменшується. Це можна пояснити тим, що взимку менше будинків пропонується на продаж. Весняний період вважається піковим не тільки через більшу пропозицію житла, але й через більшу кількість покупців. Завдяки конкуренції за будинки навесні, ціни часто піднімаються, але

влітку вони починають нормалізуватися. Це пояснює, чому травень, червень, липень і серпень є найбільш активними місяцями для придбання житла [68].

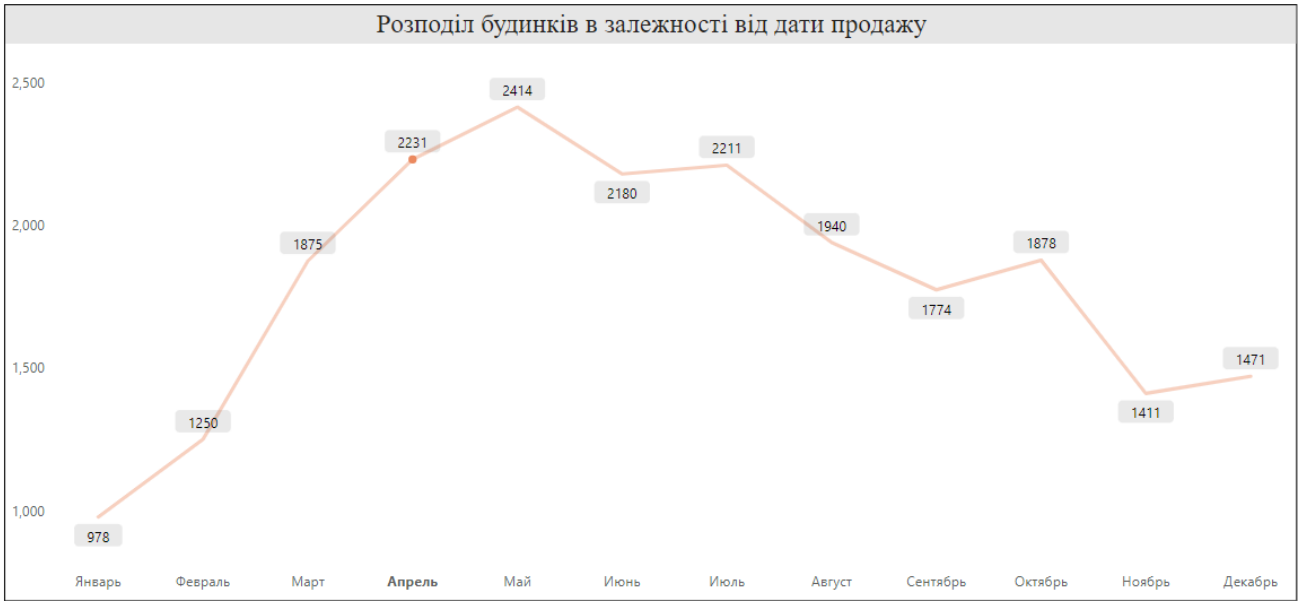


Рисунок 3.13 – Розподіл будинків в залежності від дати продажу

Наступним важним етапом є аналіз розподілу цін на будинки. На діаграмі 3.14 представлені три гістограми, що відображають розподіл середньої ціни на будинки за такими характеристиками: індекс стану будинку, індекс виду будинку та індекс відповідності будівельному та дизайнерському рівням.

Загалом видно, що найвищі вартості спостерігаються у будинків з індексом стану 5 (526 000 доларів). Це логічно, оскільки цей індекс вказує на високу якість обслуговування та оновлення будинку. Будинки з індексами 3 та 4 мають приблизно однакову середню ціну. Найнижчі вартості спостерігаються у будинків з індексами 1 та 2, що свідчить про потребу у значних ремонтах та вказує на погіршення стану.

При аналізі розподілу цін за індексом виду будинку помітно, що найбільшу середню вартість мають будинки з індексом 4 (1 185 000 доларів). Значно нижча середня ціна спостерігається у будинків з індексом 3 (802 500 доларів), а найнижча вартість — у будинків з індексом 0 (432 500 доларів).

На основі результатів розподілу цін за будівельним та дизайнерським індексами можна зробити наступні висновки. Середньою вартістю найбільш цінних будинків є ті, що мають індекс 13, що свідчить про їх вишуканість. Також можна відзначити, що будинки з індексами 11 та 12 є достатньо дорогими (вартістю 1 817 500 доларів та 1 284 000 доларів відповідно). Найнижчу вартість мають будинки з індексами 1.5, що виправдано, оскільки такі будинки зазвичай мають недостатній ремонт та просту конструкцію.

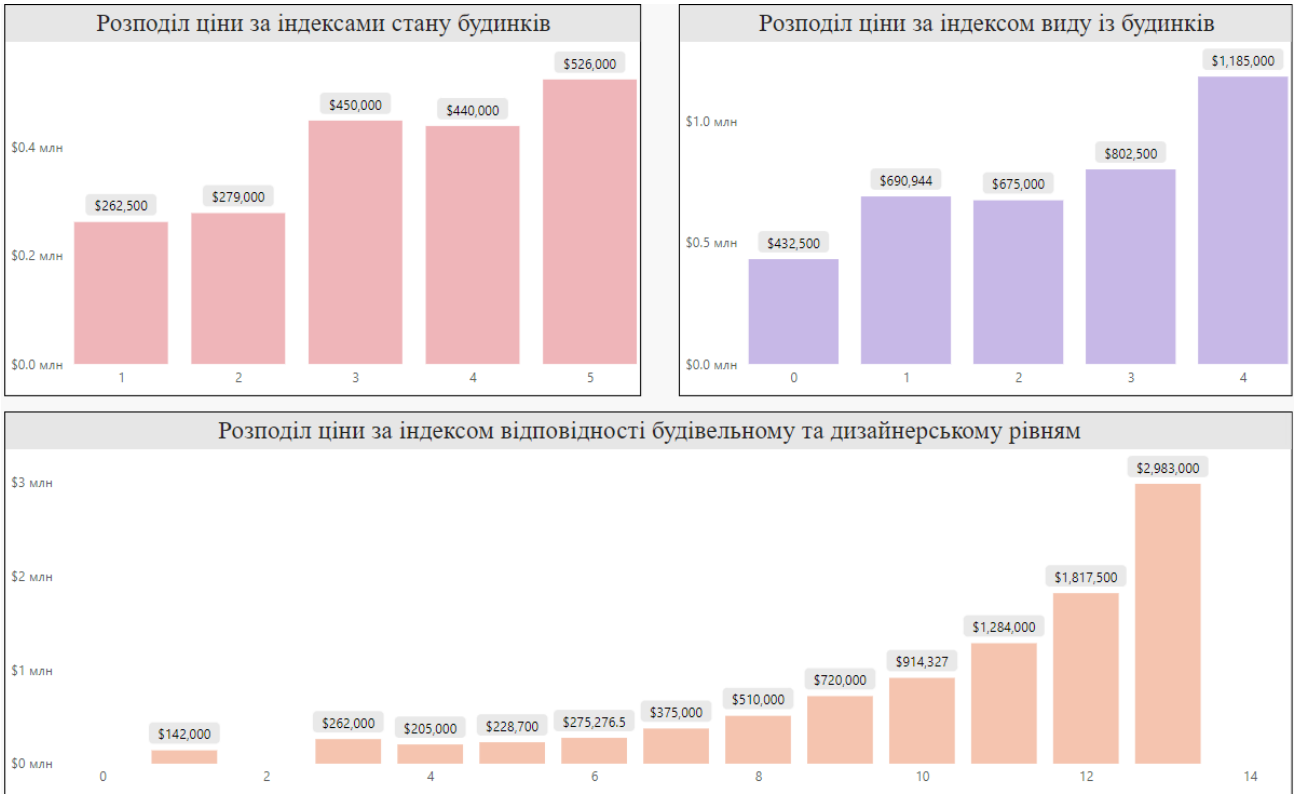


Рисунок 3.14 – Розподіл ціни будинків за індексами

На рисунку 3.15 можна спостерігати середню вартість будинків у різних містах. Зробивши аналіз, ми приходимо до висновку, що Медіна має найвищу середню ціну на будинки (1892 500 доларів), хоча, як було показано раніше, в цьому місті є найменша кількість будинків - всього 50. Цю високу ціну можна пояснити тим, що місто оточене озером Вашингтон з півночі, заходу та півдня. Берегова лінія Медіни є однією з найкращих набережних в окрузі Кінг. Крім

того, дорогі будинки також знаходяться в Мерсер Айленд (середня ціна - 993 750 доларів).

Це пояснюється його вигідним розташуванням на окремому острові, що оточений озером Вашингтон з усіх сторін. Мерсер Айленд відомий своїм достатком і знаменитими мешканцями, і є одним з найбагатших міст штату Вашингтон за доходом на душу населення. За найнижчими цінами на будинки можна знайти у містах Кент, Енумклау, Аубурн і Федерал Вей (приблизно 200 000 доларів), які знаходяться поруч одне з одним, згідно з картою.

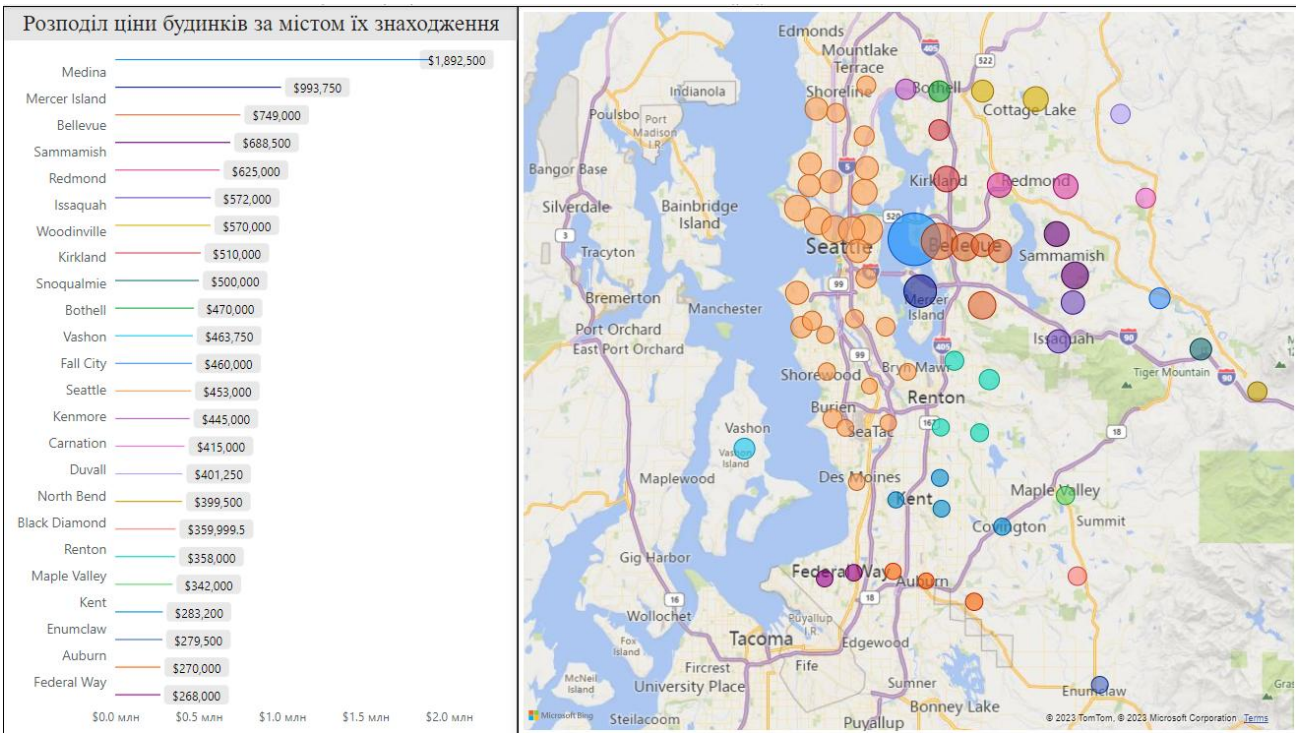


Рисунок 3.15 – Розподіл ціни будинків за містом їх знаходження

З аналізу рисунку 3.16 можна зробити висновок, що найвищою середньою вартістю володіють будинки, побудовані менше 10 років тому (550 000 доларів). Також можна побачити, що значну цінність мають будинки з віком від 91 до 100 років (540 000 доларів), а також ті, яким вже понад 100 років (527 500 доларів).

Це може бути пов'язано з тим, що нові будинки мають сучасний дизайн та використовуються новітні будівельні матеріали, а старі будинки можуть мати

вишуканий архітектурний стиль та історичну цінність. Найменшою вартістю володіють будинки з віком від 51 до 60 та від 71 до 80 років (380 000 доларів).

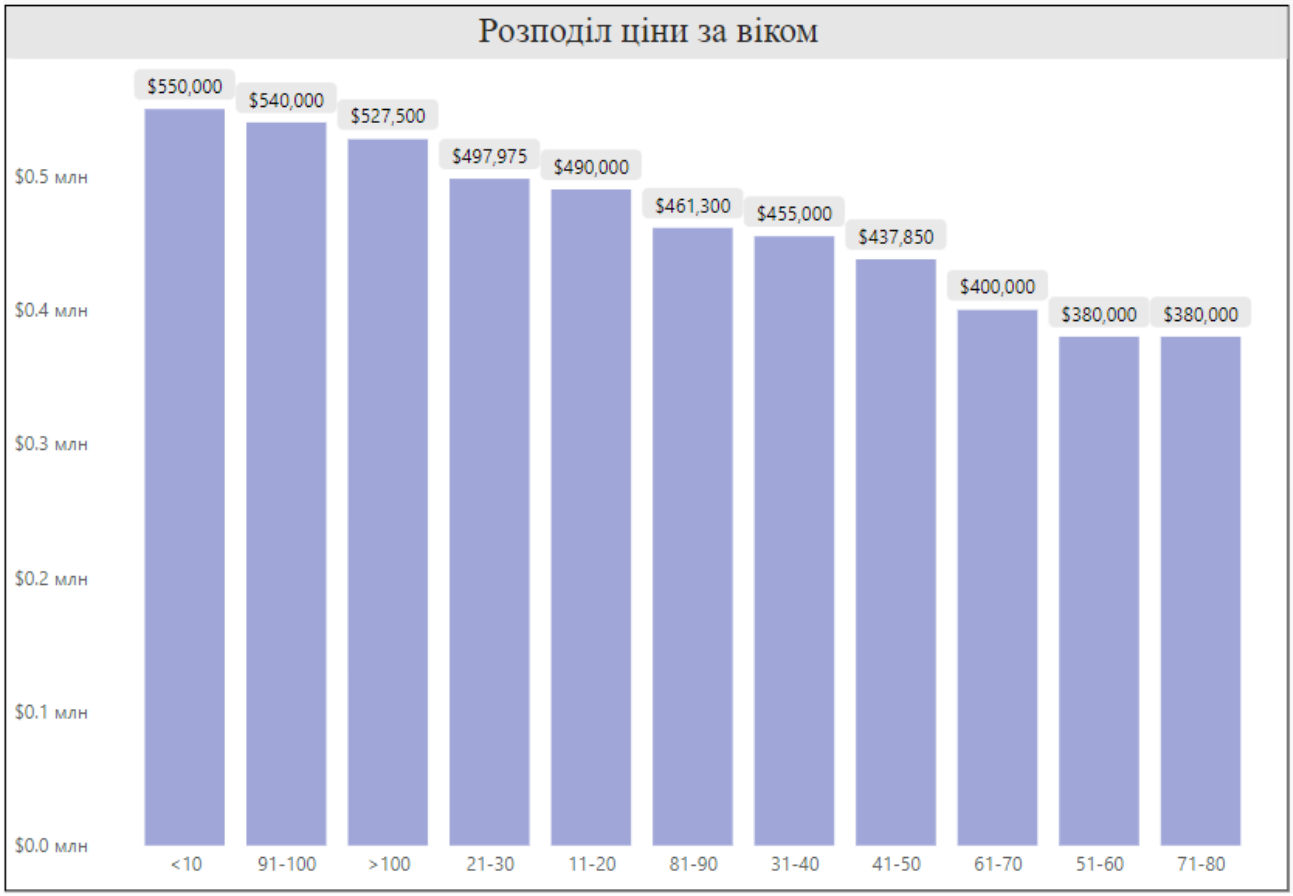


Рисунок 3.16 – Розподіл ціни будинків за віком

Також, слід провести аналіз залежності ціни від житлової площі будинку, результати якого можна знайти на рис. 3.17. З огляду на житлову площу, можна зробити висновок, що в середньому будинки з площею понад 8 000 квадратних футів (5 110 800 доларів) мають найвищу вартість, тоді як будинки з площею менше 1 000 квадратних футів (285 000 доларів) мають найнижчу вартість.

Подібну тенденцію можна спостерігати і в розподілі ціни залежно від площі ділянки. В середньому, найбільші вартості спостерігаються для будинків з площею понад 400 000 квадратних футів (722 500 доларів), тоді як найнижчі вартості зафіксовані для будинків з площею до 10 000 квадратних футів (431 000 доларів). Варто відмітити, що більшість груп площ мають симетричний розподіл.

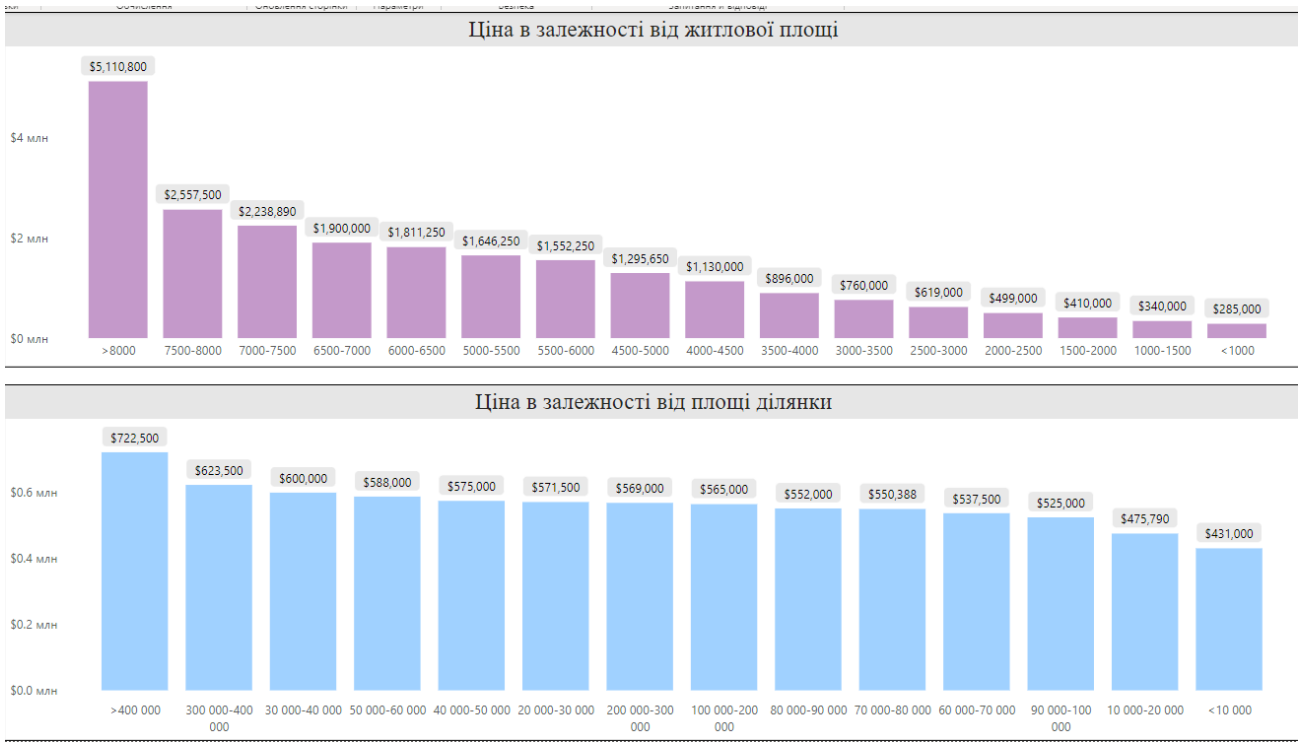


Рисунок 3.17 – Розподіл ціни будинків в залежності від площі

Варто проаналізувати розподіл цін в залежності від року будівництва. Згідно з діаграмою 3.18, можна помітити, що середня ціна на будинки поступово зменшується, зокрема найнижчою виявилася ціна на будинки, побудовані в 1943 році (287 450 доларів). Однак, варто зазначити, що ціна на будинки, побудовані після 1943 року, зростає з кожним наступним роком будівництва.

Крім того, важливо відзначити, що в період з 2007 по 2009 рік середня вартість будинків помітно зменшилася, що можна пояснити світовою фінансовою кризою, що виникла у 2007-2008 роках.

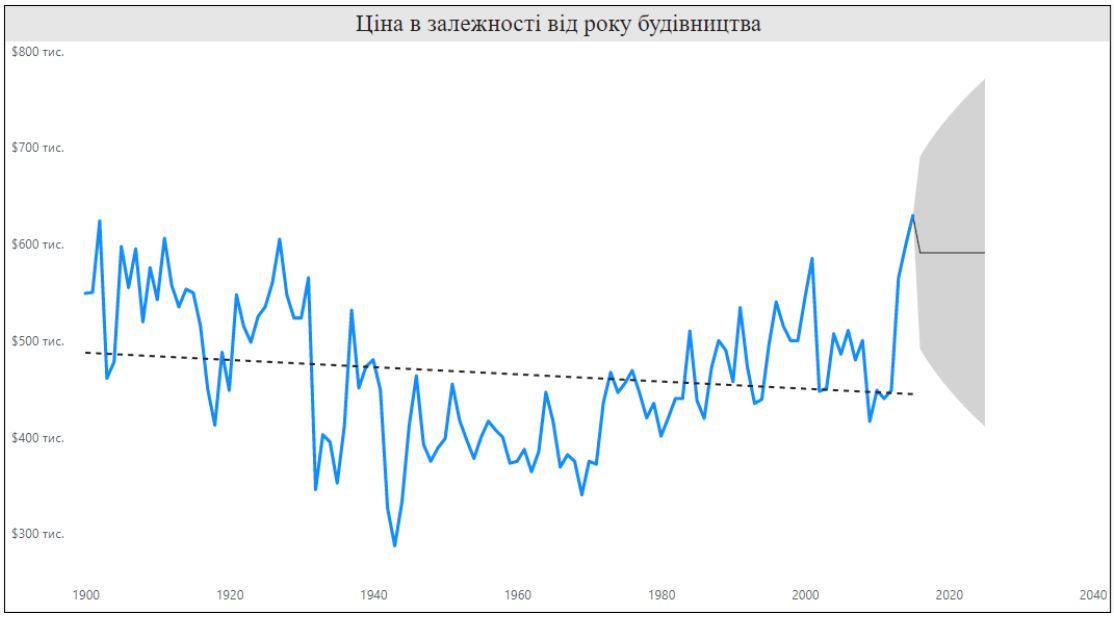


Рисунок 3.18 – Розподіл ціни будинків в залежності від року будівництва

3.2 Моделювання ціни на нерухомість методом лінійної моделі

Перш ніж приступити безпосередньо до моделювання, потрібно виконати підготовчі заходи. По-перше, у середовищі програмування R необхідно завантажити дані та всі необхідні бібліотеки для роботи. Крім того, можна вилучити з даних змінну ID, оскільки вона є лише ідентифікатором спостереження і не впливає на ціну. Наступним важливим кроком є виявлення викидів у цінах. Викиди - це незвичайні значення в наборі даних, які можуть спотворити статистичний аналіз. Існує багато методів для виявлення потенційних викидів, але я вирішила скористатися методом інтерквартильного діапазону (IQR) у своїй роботі. Будь-який набір даних можна описати за допомогою його п'яти числових характеристик. Ці п'ять чисел, які надають необхідну інформацію для пошуку відхилень, включають:

– мінімальне або найменше значення набору даних;

– перший квартиль (Q1), який визначається як середнє між найменшим числом (мінімумом) та медіаною набору даних. Його також

називають нижнім або 25-м емпіричним квартилем, оскільки 25% даних знаходяться нижче цієї точки;

– другий квартиль (Q2), який є медіаною набору даних і розділяє його на дві рівні частини, тобто 50% даних знаходяться нижче цієї точки;

– третій квартиль (Q3), який є середнім значенням між медіаною та найвищим значенням (максимумом) набору даних. Його також називають верхнім або 75-м емпіричним квартилем, оскільки 75% даних знаходяться вище цієї точки;

– максимальне або найвище значення набору даних.

Інтерквартильний метод використовується для виявлення викидів у числових наборах даних. Цей метод включає наступні кроки:

1. Знаходження першого квартиля (Q1).
2. Знаходження третього квартиля (Q3).
3. Обчислення міжквартильного розмаху (IQR) за допомогою формули (3.1):

$$IQR = Q3 - Q1 \tag{3.1}$$

4. Визначення нормального діапазону даних з нижньою та верхньою межами, використовуючи формули (3.2). Константа 1,5 використовується для відрізнення викидів.

$$Q1 - 1{,}5 * IQR, Q3 + 1{,}5 * IQR \tag{3.2}$$

Для подальшого аналізу слід вилучити будь-яку точку, що виходить за межі цього діапазону [71]. Точки на графіку, які вважаються можливими викидами, представлені як спостереження. Згідно з цим, на рис. 3.19 можна виділити кілька потенційних викидів.

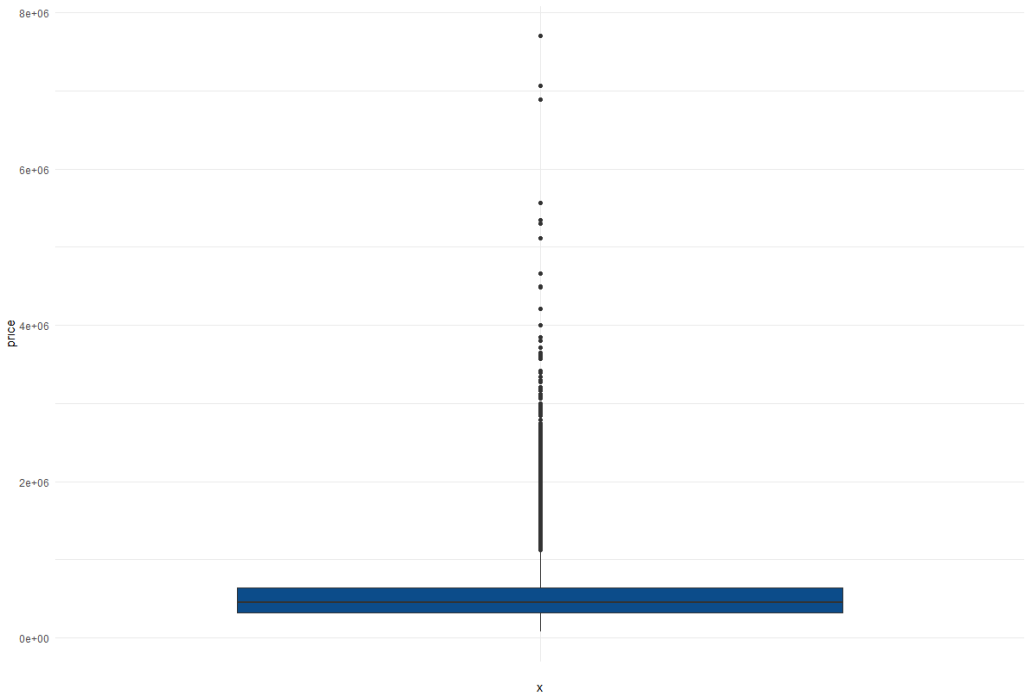


Рисунок 3.19 – Графік викидів ціни

Після застосування інтерквартильного методу виявлення викидів до даних, залишилося 20 467 спостережень, що становить 1 146 значень, визначених як викиди.

Згідно з кореляційною матрицею, можна спостерігати кореляцію 1 між показниками "yr_built" та "age". Це може вплинути на результати моделі та призвести до обманливих висновків. Тому краще видалити один з цих показників. Оскільки ці показники містять однакову інформацію, не має значення, який з них буде видалено. Я вирішила прибрати "yr_built", оскільки, як показано у першому розділі цієї роботи, вік будинку відіграє важливу роль.

Крім того, логічний зв'язок між полями значень року реновації (yr_renovated) та року ремонту (repair_age) дорівнює 0,92, що є досить високим значенням. Тому потрібно провести аналіз на наявність мультиколінеарності та визначити значення показника VIF. Цей показник вимірює, наскільки висока дисперсія коефіцієнта регресії через наявність мультиколінеарності у моделі.

Зазвичай значення VIF, що перевищує 5 або 10, вказує на проблематичний рівень колінеарності.

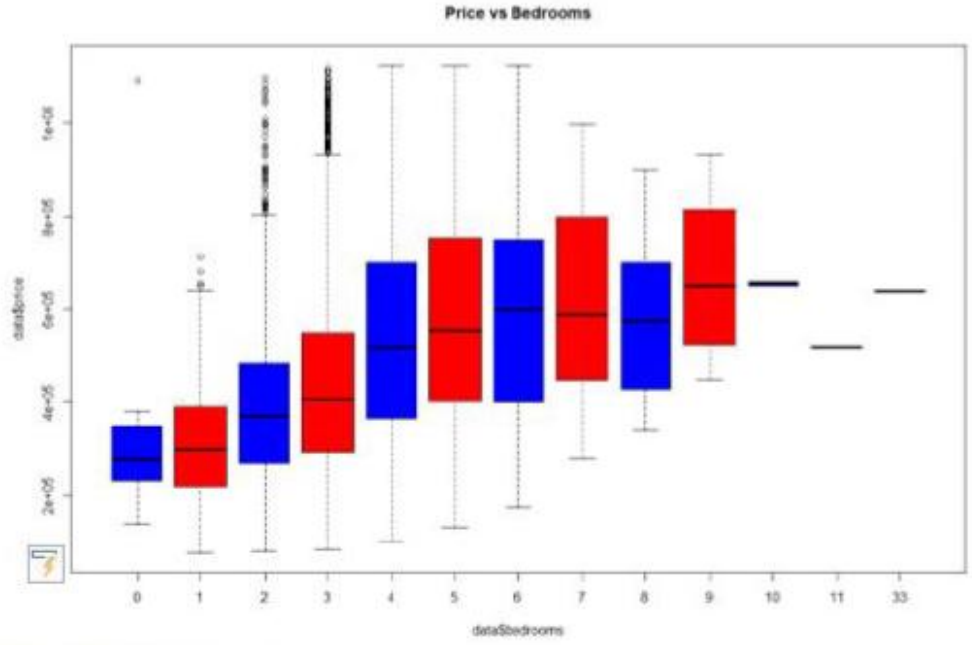
Зглядом на графік 3.20 можна зробити висновок, що значення VIF перевищує 5, що означає, що доцільно вилучити один з показників. Так як зв'язок між ціною на нерухомість та роком ремонту є сильнішим (0.13), то ми вирішуємо видалити показник repair_age.

```
Call:
lm(formula = price ~ yr_renovated + repair_age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-479224 -160209  -38556  126444  653944

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  473556.070   1477.966 320.411  < 2e-16 ***
yr_renovated    -11.232      9.796  -1.147    0.252
repair_age     2002.039    315.465   6.346 2.25e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 207400 on 20464 degrees of freedom
Multiple R-squared:  0.00894,   Adjusted R-squared:  0.008843
F-statistic:  92.3 on 2 and 20464 DF,  p-value: < 2.2e-16

> vif(model1)
yr_renovated   repair_age
    6.547052     6.547052
```

Рисунок 3.20 – Перевірка на мультиколінеарність

Давайте проаналізуємо взаємозв'язок між ціною та показниками за допомогою бокс-плотів. Бокс-плот є стандартизованим способом візуалізації розподілу даних, як я було згадано раніше. На рис. 3.21 зображено бокс-плот, що показує зв'язок між ціною та кількістю спалень.

Згідно отриманих результатів, варто перевірити будинки з 11 та 33 спальнями, оскільки їх ціна здається занадто низькою. Будинок з 33 спальнями має маленьку житлову площу для такої кількості спалень і лише 1,75 ванної кімнати. Також середня ціна будинку з 11 спальнями також здається занадто низькою. Ці будинки вважаються аномальними спостереженнями, тому краще їх видалити. Важливо відзначити, що кількість спалень не має лінійного зв'язку з ціною, але є скоріше категоріальним показником. В середовищі програмування R категоріальні показники необхідно перетворити на фактор.




```
> print(subset(data, data$bedrooms > 10))
# A tibble: 2 x 17
   price  date bedrooms bathrooms sqft_living sqft_lot floors waterfront  view condition grade sqft_above sqft_basement
   <dbl> <dbl>    <dbl>     <dbl>       <dbl>    <dbl>  <dbl>      <dbl> <dbl>     <dbl> <dbl>     <dbl>        <dbl>
1 520000 43333       11      3          3000     4960      2          0     0         3     7      2400          600
2 640000 43276       33      1.75       1620     6000      1          0     0         5     7      1040          580
# ... with 4 more variables: age <dbl>, zipcode <dbl>, lat <dbl>, long <dbl>
```

Рисунок 3.21 – Графік залежності між ціною та спальними кімнатами

На рисунку 3.22 показана діаграма розкиду (бокс-плот) для залежності між ціною та кількістю ванних кімнат. Між цими двома змінними спостерігається лінійний зв'язок.

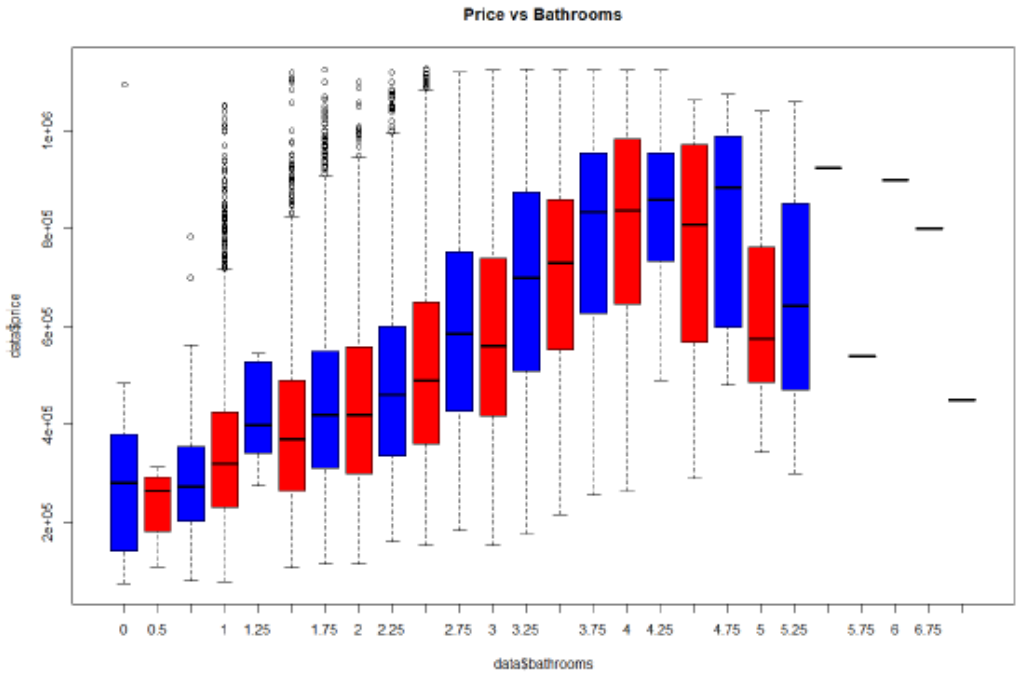


Рисунок 3.22 – Аналіз залежності між ціною та ванними кімнатами

На рисунку 3.23 зображений бокс-плот, що відображає взаємозв'язок між ціною та кількістю поверхів. Виявляється, що кількість поверхів не впливає лінійно на ціну, але скоріш є категоріальним показником. Тому ми перетворюємо його на фактор.

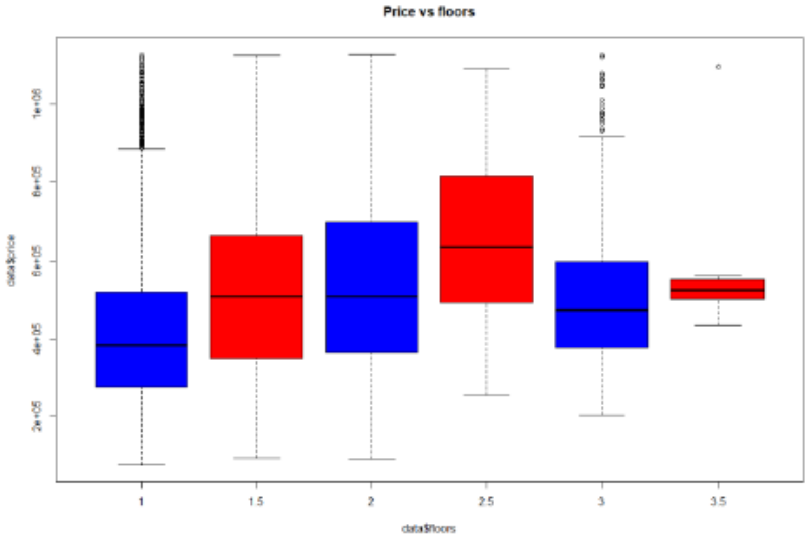


Рисунок 3.23 – Аналіз залежності між ціною та поверхами

Ми можемо перетворити вихід на набережну, який представлений на рис. 3.24, у фактор, оскільки він не має лінійного зв'язку з ціною і є скоріш категорійним показником.

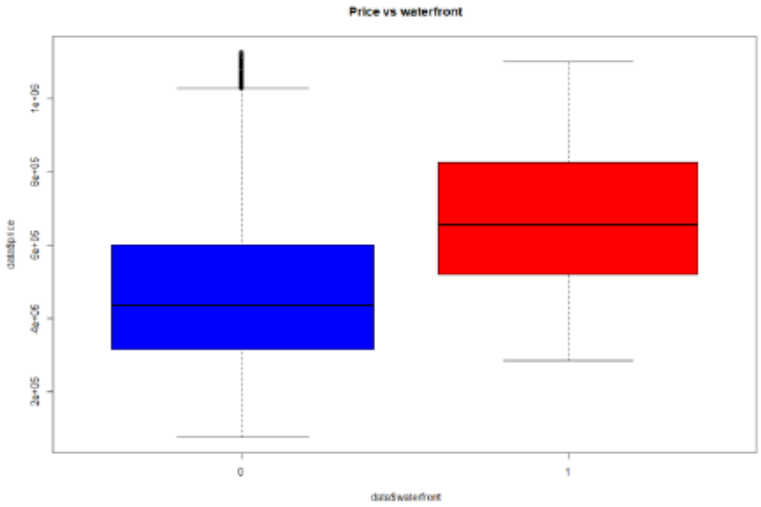


Рисунок 3.24 – Аналіз залежності між ціною та виходом на набережну

На рисунку 3.25 представлений бокс-плот, що ілюструє залежність між ціною та типом будинку. Тип будинку не має лінійної залежності від ціни; скоріше, він є категоріальним показником. Тому ми перетворюємо його на фактор.

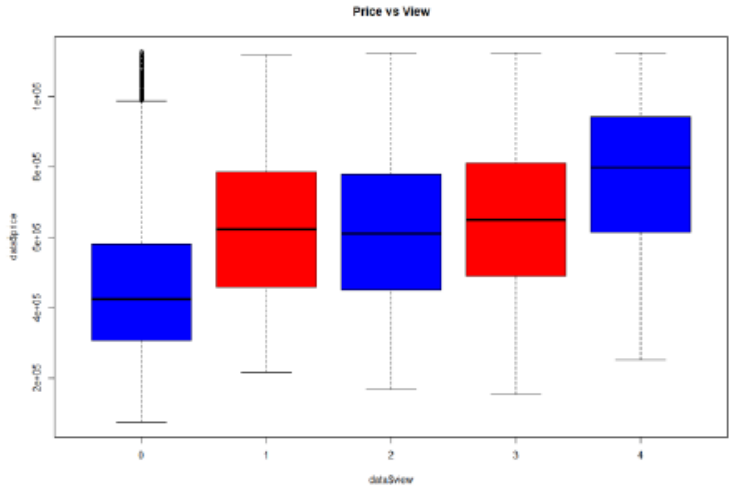


Рисунок 3.25 – Аналіз залежності між ціною та видом з будинку

Рисунок 3.26 показує бокс-плот, який відображає залежність між ціною та станом будинку. Стан будинку не демонструє лінійного зв'язку з ціною, але він

може бути більше розглянутий як категоріальний показник. Тому, ми перетворимо його на фактор.

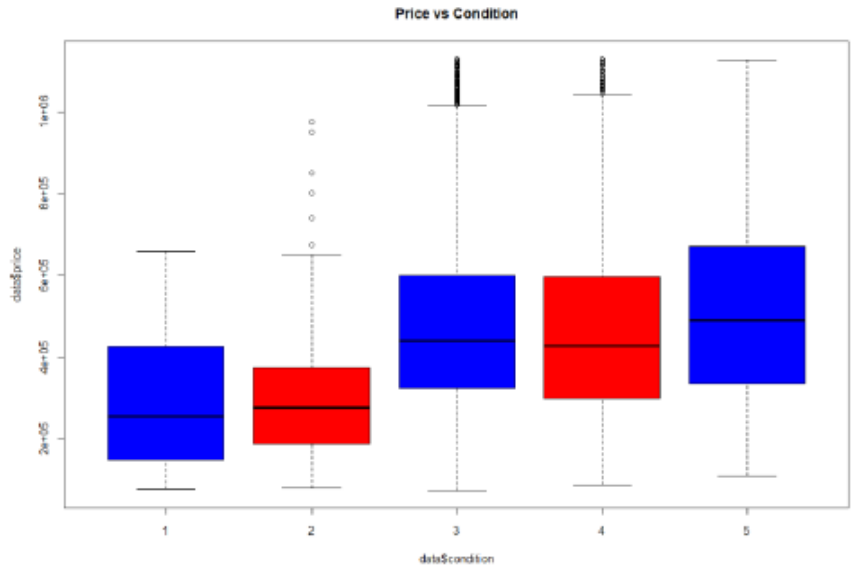


Рисунок 3.26 – Аналіз залежності між ціною та станом будинку

На рис. 3.27 зображений бокс-плот, що демонструє залежність між ціною та відповідністю будівельному та дизайнерському рівням. Ці дві змінні мають лінійну залежність між собою.

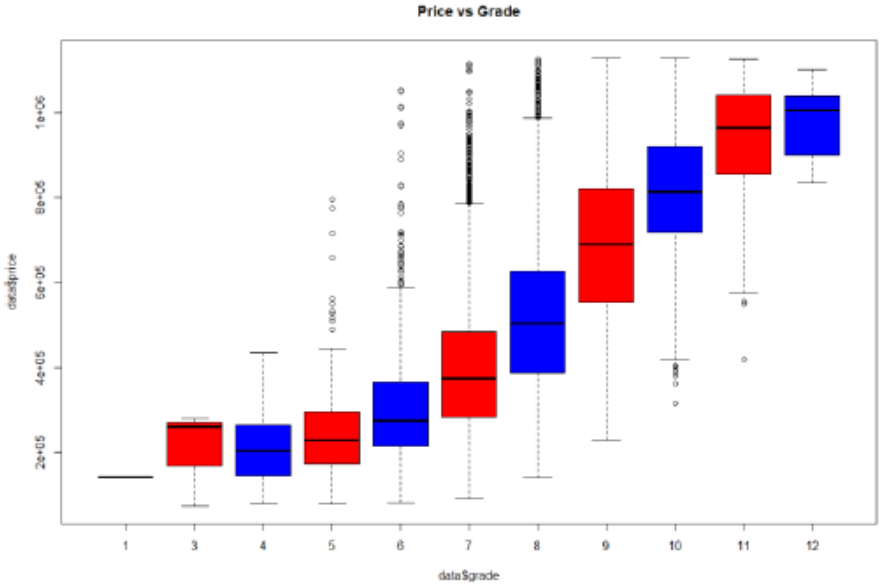


Рисунок 3.27 – Залежності між ціною та відповідністю будівельному та дизайнерському рівням

Згідно отриманих даних, здавалося, що більшість будинків не мають підвалу, тому була проведена перевірка. Було встановлено, що 12 702 будинків не мають підвалу. Оскільки кількість будинків без підвалу перевищує половину, можна поділити вибірку на дві категорії - 0 і 1. Будинки без підвалу будуть мати значення 0, тоді як будинки з підвалом отримають значення 1, як показано на рис. 3.28.

Таким чином, цей показник перетворився на категоріальну змінну, яку ми можемо розглядати як фактор.

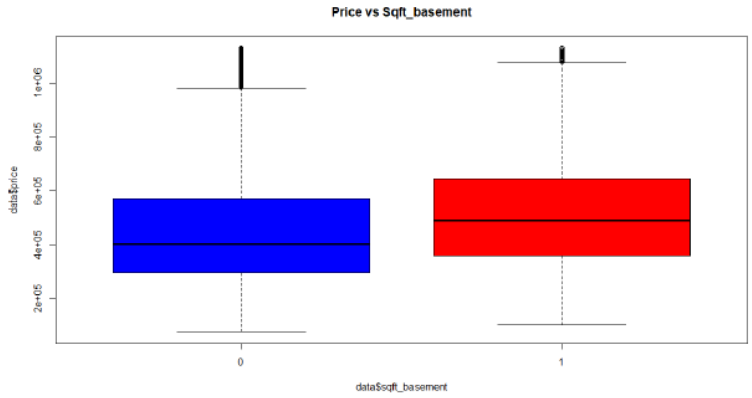


Рисунок 3.28 – Аналіз залежності між ціною та квадратним футом внутрішнього простору житла

За даними можна зробити висновок, що більшість будинків не пройшла ремонт, тому була проведена докладна перевірка. Було встановлено, що 19 700 будинків не мають свіжого ремонту. Оскільки це значно більше половини вибірки, ми можемо розділити дані на дві категорії, позначивши будинки без ремонту значенням 0, а ті, що мають ремонт, - значенням 1. Цю інформацію можна побачити на рисунку 3.29. Таким чином, цей показник стає категоріальним, і ми перетворюємо його на фактор.

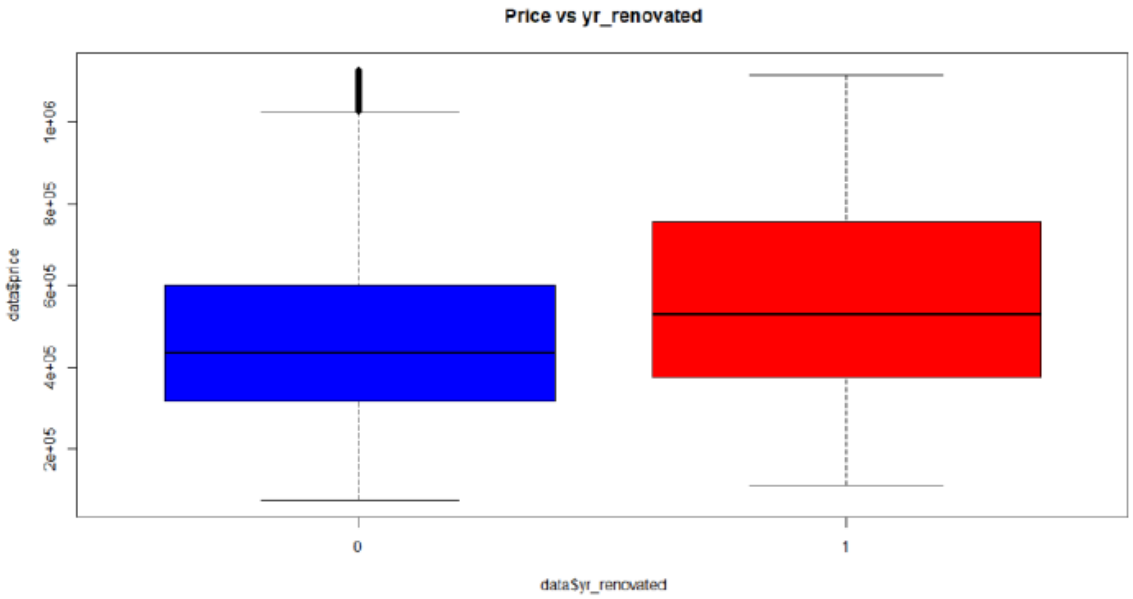


Рисунок 3.29 – Аналіз залежності ціною та роком останнього ремонту

Рисунок 3.30 відображає бокс-плот, який показує зв'язок між ціною та поштовим кодом будинку. Стан будинку не має прямої лінійної залежності від ціни, але може бути розглянутий як категоріальний показник. Тому ми перетворюємо його в фактор.

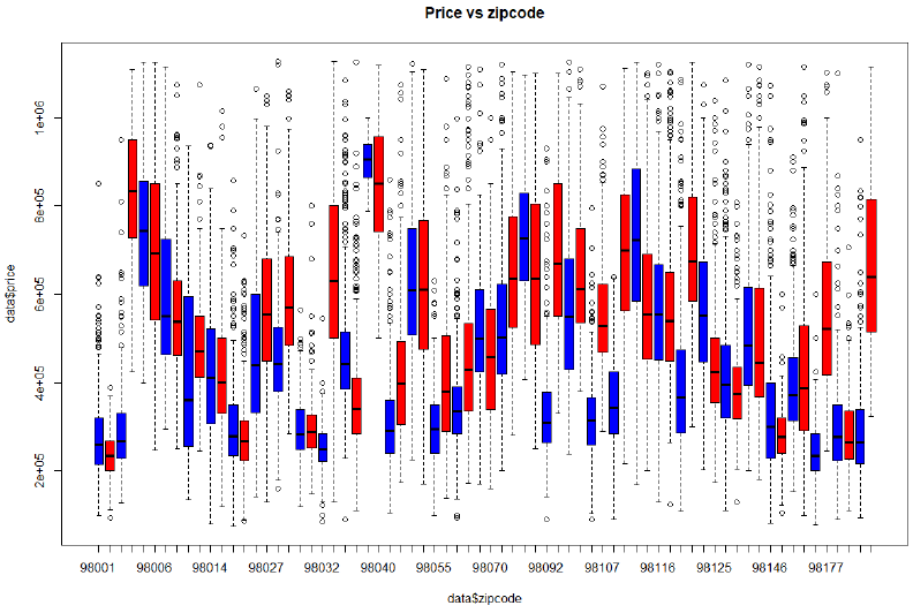


Рисунок 3.30 – Аналіз залежності між ціною та поштовим кодом

Перед початком моделювання, спочатку необхідно провести нормалізацію даних. Метою нормалізації є перетворення значень числових стовпців в наборі

даних таким чином, щоб вони знаходилися на загальній шкалі, зберігаючи відмінності у діапазонах значень та інформацію. У моїй роботі я здійснив нормалізацію лише незалежних змінних, оскільки ціна, як залежна змінна використається для проведення прогнозування, і її значення має бути у початкових одиницях виміру.

Також, розбиття даних на навчальний та тестовий датасети є важливою частиною перевірки моделей. Зазвичай, більша частина даних використовується для навчання, а менша частина - для тестування. Ефективність моделі перевіряється саме на тестовому наборі даних. Я розділив свій набір даних, що 70% значень потрапити у вибірку для навчального датасету (14326), а 30% (6139) були залишені для тестування.

Початковою моделлю, що була побудована, є множинна лінійна регресія. Спочатку на навчальному наборі даних була побудована лінійна модель з усіма 18 змінними. Після виконання відповідного R-коду було отримано результат, який можна побачити на рис. Б.1. З цього графіка можна зробити певні висновки.

Формула виклику - це формула R, яка використовується для розташування даних. Залишки - це різниця між фактичними спостережуваними значеннями у і значеннями, які передбачала модель. Щоб оцінити, наскільки модель відповідає даним, слід шукати симетричний розподіл між цими точками навколо середнього значення нуля. В нашому прикладі ми бачимо, що розподіл залишків є достатньо симетричним.

Коефіцієнти представляють собою ваги, які мінімізують суму квадратів помилок. В цьому розділі описано наступне:

1. Коефіцієнт-перетин (Intercept) - це значення Y, яке отримуємо, коли X дорівнює нулю. У даному випадку значення коефіцієнта-перетину дорівнює 301109,7.

2. Оцінки (Estimate) для параметрів моделі - це кількість одиниць, на яку змінюється Y, коли X змінюється на 1 одиницю.

3. Стандартна помилка розрахункових значень (Std. Error) - це міра точності оцінки коефіцієнта. Вона вимірює, наскільки точно модель оцінює невідоме значення коефіцієнта.

4. Значення коефіцієнта t - це міра віддаленості оцінки коефіцієнта від нуля в стандартних відхиленнях. У цьому прикладі значення t-статистики в основному далекі від нуля і значно перевищують стандартну помилку. Це може свідчити про наявність зв'язку між залежною та незалежними змінними.

5. P-значення - це індивідуальне значення p для кожного параметра, яке приймає або відхиляє нульову гіпотезу. Якщо значення p низьке (менше 5%), це свідчить про можливість відхилення нульової гіпотези. В нашому випадку, p-значення дуже близьке до нуля, що вказує на можливість відхилення нульової гіпотези, і підтверджує наявність залежності між незалежними величинами та ціною.

Залишкова стандартна помилка визначає середнє відхилення змінної від істинної лінії регресії. У нашому прикладі, ціна може в середньому відхилятися від справжньої лінії регресії на приблизно 85490 доларів.

Іншими словами, з урахуванням того, що середня ціна для всіх будинків становить 301109,7, залишкова стандартна помилка складає 85490, що еквівалентно 28,3% відсотковій помилці.

Множинний R-квадрат і відрегульований R-квадрат є статистиками, що вказують, наскільки добре модель відповідає фактичним даним. У нашому випадку, значення множинного R-квадрата дорівнює 0,8328 або приблизно 83% дисперсії залежної змінної можна пояснити змінними незалежних змінних. Скоригований R-квадрат має значення 0,8316, що також вказує на те, що близько 83% дисперсії залежної змінної можна пояснити змінними незалежних змінних.

F-статистика використовується для визначення наявності зв'язку між залежною та незалежними змінними. Чим значення F-статистики далі від 1, тим краще. В нашому випадку, значення F-статистики дорівнює 681,3, що є великим значенням порівняно з 1, враховуючи розмірність наших даних [72].

Маючи p-value менше 2.2e-16, ми можемо відхилити нульову гіпотезу, що дозволяє зробити висновок про існування залежності між ціною та незалежними змінними. Слідуючим кроком буде перевірка на тестовому наборі. Встановлено, що R-квадрат дорівнює 0,8371, ≈84% дисперсії залежної змінної можна пояснити змінними, які не залежать від неї.

Подальше проведено обчислення точності моделі прогнозу, результати яких такі:

- Коренева середня квадратична помилка (RMSE) = 85054,44
- Середня абсолютна помилка (MAE) = 61814.5
- Відносна абсолютна помилка (MAPE) = 0,1456575
- Симетрична відносна абсолютна помилка (SMAPE) = 0,143053
- Середня абсолютна відносна помилка (MASE) = 0,2652008

Зазвичай, для оцінки ефективності моделі найкращими показниками є RMSE та MAPE. RMSE вказує на відхилення середніх спрогнозованих значень від тестових на 85054,44 доларів.

Складно оцінити це як добре чи погано, не дивлячись на те, що викиди були прибрані, ціна коливається від мільйонів доларів до тисяч. В такій ситуації для інтерпретації результатів допомагає показник MAPE, який вказує, що прогнозовані значення в середньому відхиляються від тестових на 14,5%, що є задовільним показником.

Наступним кроком було створення поетапної регресії, яка включає ітеративне додавання та видалення незалежних змінних у моделі з метою знаходження набору змінних, що приводить до найефективнішої моделі, тобто моделі, яка мінімізує помилку передбачення. Було створено три моделі:

включення змінних, виключення змінних та комбінація обох підходів. Після запуску відповідних кодів на рисунку 3.31 видно, що кожна зі стратегій дала такий самий результат, як і повна регресія. Жодна з незалежних змінних не була вилучена.

```
Start:  AIC=325480.5
price ~ date + bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + sqft_above +
    sqft_basement + yr_renovated + age + zipcode + lat + long

               Df  Sum of Sq        RSS    AIC
<none>                       1.0393e+14 325481
- long          1 1.9249e+10 1.0395e+14 325481
- sqft_basement  1 4.1527e+10 1.0397e+14 325484
- lat           1 1.3908e+11 1.0407e+14 325498
- bathrooms     1 3.0692e+11 1.0424e+14 325521
- bedrooms     10 5.8345e+11 1.0451e+14 325541
- age           1 5.6856e+11 1.0450e+14 325557
- yr_renovated  1 5.7961e+11 1.0451e+14 325558
- waterfront    1 6.7887e+11 1.0461e+14 325572
- floors        5 9.1491e+11 1.0485e+14 325596
- date          1 1.3280e+12 1.0526e+14 325660
- sqft_above    1 1.4648e+12 1.0539e+14 325679
- sqft_lot      1 1.7264e+12 1.0566e+14 325715
- condition     4 2.7508e+12 1.0668e+14 325847
- sqft_living   1 2.8365e+12 1.0677e+14 325864
- view          4 5.9441e+12 1.0987e+14 326269
- grade         1 1.1546e+13 1.1548e+14 326988
- zipcode      69 8.8213e+13 1.9214e+14 334146

Call:
lm(formula = price ~ date + bedrooms + bathrooms + sqft_living +
    sqft_lot + floors + waterfront + view + condition + grade +
    sqft_above + sqft_basement + yr_renovated + age + zipcode +
    lat + long, data = train)
```

Рисунок 3.31 – Результат лінійної поступової регресійної моделі

Також було створено модель, з якої я особисто виключив незначні показники ціни, такі як дата (0) і довгота (0,02). Результат моделювання можна побачити на рис. Б.1. У цьому прикладі скоригований коефіцієнт детермінації R2 дорівнює 0,8295, що означає, що приблизно 83% дисперсії залежної змінної можна пояснити незалежними змінними.

Після проведення ефективності перевірки на тестовому наборі, можна зробити висновок, що коефіцієнт детермінації R-квадрат дорівнює 0,831, або приблизно 83%. Це означає, що близько 83% дисперсії, виявленої у залежній змінній, можна пояснити за допомогою предикторної змінної. Наступним кроком було розрахувати точність моделі прогнозу, і отримані результати такі:

- Корінь середньоквадратичної помилки (RMSE) = 85729,96
- Середня абсолютна помилка (MAE) = 62339,5
- Середня відносна абсолютна помилка (MAPE) = 0,1467838

- Симетрична середня відносна абсолютна помилка (SMAPE) = 0,1440709

- Масштабована абсолютна помилка (MASE) = 0,2674532

Значення RMSE вказує на відхилення середніх спрогнозованих значень від тестових на 85729,96 доларів. Показник MAPE показує, що у середньому прогнозовані значення відхиляються від експериментальних на 14,7%.

Останнім кроком буде порівняння двох моделей за допомогою показників, таких як R-квадрат, RMSE та MAPE, які можна побачити на рис. 3.32. Зауважимо, що значення R-квадрат для повної регресії є кращим. Тому значення саме цієї моделі (серед лінійних) будуть порівнюватися наприкінці з результатами інших моделей.

	R-squared	RMSE	MAPE
Full Regression	0.8337148	85054.44	0.1456575
Regression NK	0.8310677	85729.96	0.1467838

Рисунок 3.32 – Оцінки ефективності лінійних регресійних моделей

3.3 Моделювання цін на нерухомість методом регуляризації

Для побудови наступної моделі використано регуляризовану регресію. Спочатку було виконано стандартизацію X та Y. Далі вибірка була розбита на навчальний набір (70%) та тестовий набір (30%) для обох змінних X та Y. Кількість елементів у навчальному наборі для X становить 243542, а для Y - 14326. У тестовому наборі для X міститься 104363 елементи, а для Y - 6139 елементів.

Параметр альфа відіграє важливу роль у регуляризації. Якщо альфа дорівнює 0, то отримуємо рідж-регресію. При альфа рівному 1, отримуємо ласо-регресію, а якщо альфа знаходиться в межах від 0 до 1, то отримуємо еластичну сітку.

Я вирішив почати регуляризацію, використовуючи рідж регресію для обробки даних. На рисунку 3.33 можна побачити, як коефіцієнти моделі

покараються в залежності від значення показника λ. Можна зробити висновок, що при λ = 0 немає жодного ефекту, і наша цільова функція стає такою ж, як звичайна цільова функція регресії. З іншого боку, чим більше значення параметра λ, тим більший штраф накладається на коефіцієнти.

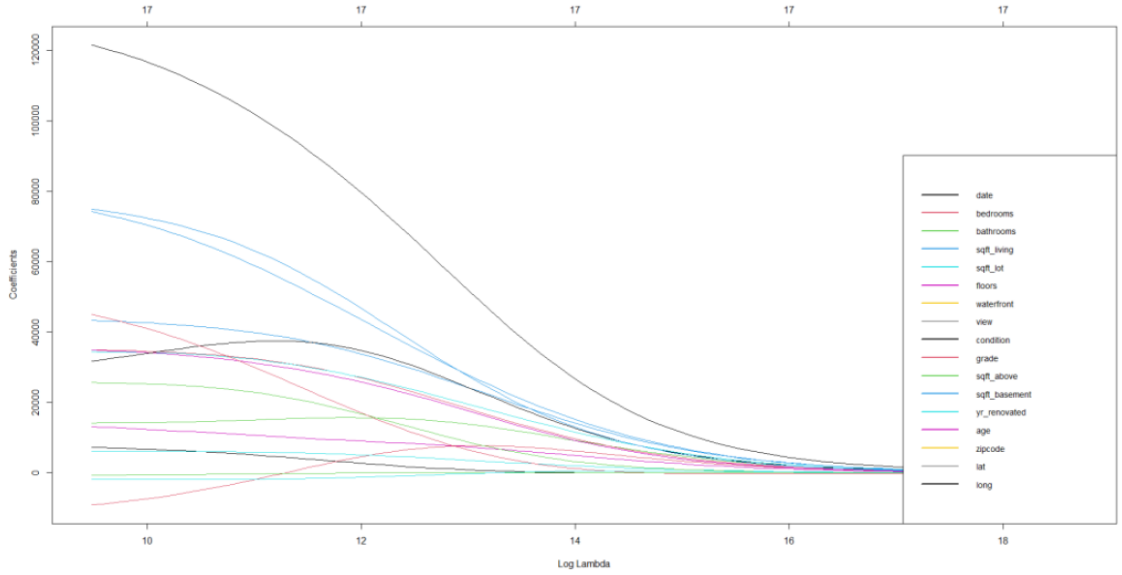


Рисунок 3.33 – Коефіцієнти рідж регресії при збільшенні значення λ

Для визначення оптимального значення лямбда для моделі необхідно виконати перехресну перевірку. На рисунку 3.34 показана 10-кратна середньоквадратична помилка CV (MSE) для різних значень лямбда. Зробивши висновок, що ми обмежуємо коефіцієнти за умовою log (λ) ≥ 0, помилка MSE значно зростає. Цифри у верхній частині графіку (17) показують кількість змінних у моделі. Змінні до нуля не штрафуються та залишаються в моделі.

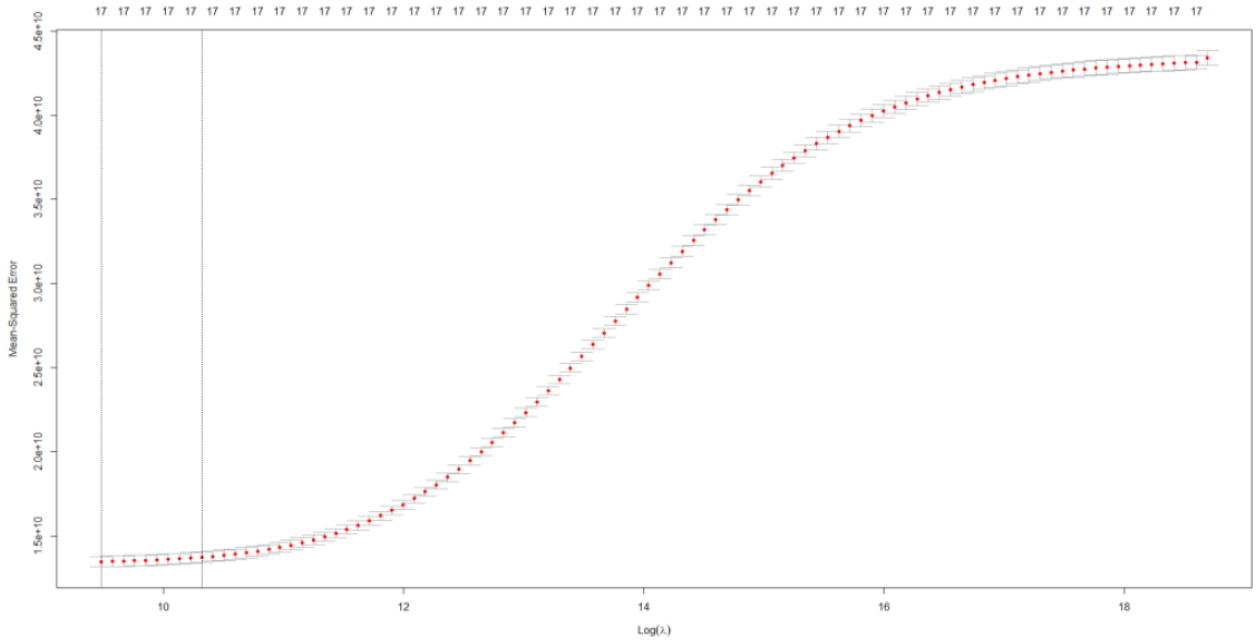


Рисунок 3.34 – Результат перехресної перевірки для визначення оптимального значення λ

Перші та другі вертикальні пунктирні лінії показують значення λ з найменшим та найбільшим значенням MSE в межах однієї стандартної похибки мінімального значення MSE. Отримані значення наступні:

- Мінімальний MSE = 13,470,619,861, λ = 13,110.69
- Стандартна похибка мінімального MSE = 13,638,548,675, λ = 25,145.12

На основі цих значень і даних рисунка 3.35 можна зробити висновок, що краще використовувати λ, рівне одній стандартній похибці, оскільки λ мінімального MSE майже не впливає на змінні.

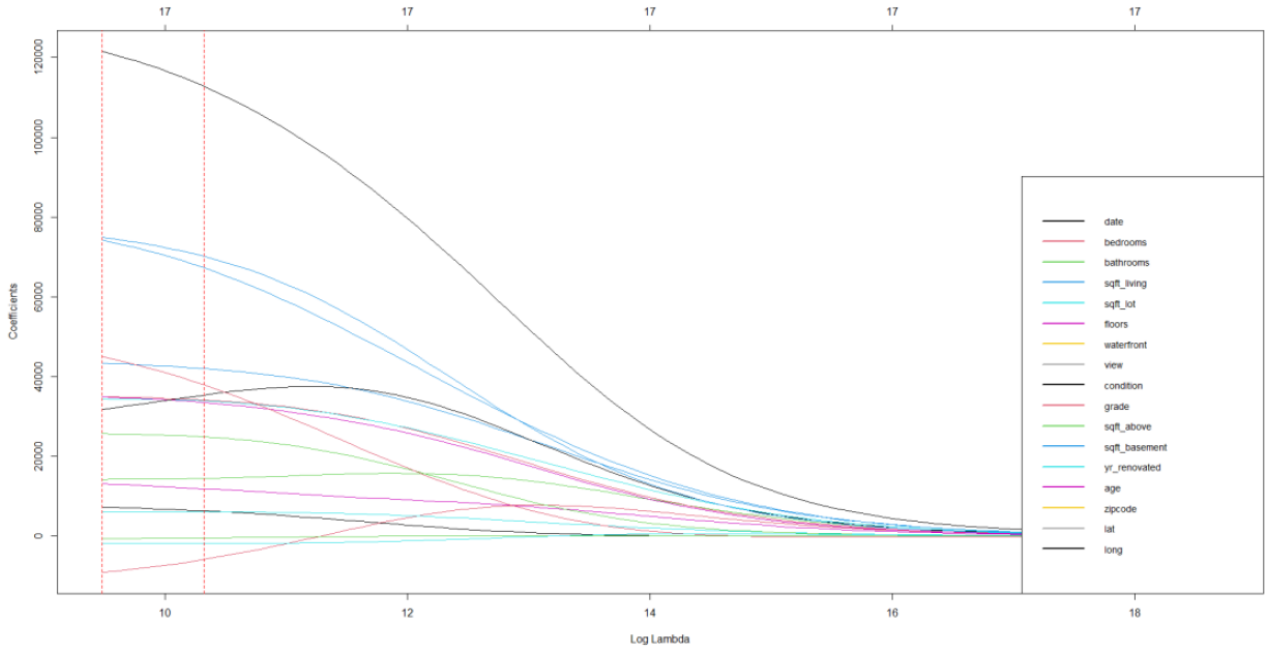


Рисунок 3.35 – Оптимальне значення λ

Наступним кроком було побудування моделі лінійної регресії з використанням параметрів альфа=0 та лямбда рівної одній стандартній похибці. З аналізу графіка 3.36 можна зробити висновок, що значенням дисперсії 68,65%, яке спостерігається у залежній змінній, можна пояснити змінними незалежних змінних.

```
Call:  glmnet(x = x_train, y = y_train, alpha = 0, lambda = ridge$lambda.1se,      standardize = TRUE)

  Df  %Dev Lambda
1 17 68.65  25150
```

Рисунок 3.36 – Результат рідж регресії

Останнім етапом є перевірка ефективності методу регуляризації на новому наборі даних. Після розрахунків було виміряно середньоквадратичну помилку (MSE) та коефіцієнт детермінації (R^2) для моделі. Отримані значення такі: MSE = 13512244229, R^2 = 0,691. Як наступна модель з регуляризацією була обрана лассо-регресія.

Штраф за ласо фактично змушує коефіцієнти наближатися до нуля, як показано на рисунку 3.37. Таким чином, лассо-модель не тільки поліпшує прогноз за допомогою регуляризації, але також автоматично вибирає змінні. На

рисунку 3.46 можна побачити, що при log(λ) = 6 всі 17 змінних є в моделі, при log(λ) = 8 залишається 16 змінних, а при log(λ) = 11 - залишається лише 3 змінні. Отже, коли набір даних має багато змінних, лассо-регресія може використовуватися для ідентифікації та виключення змінних з найбільшим сигналом.

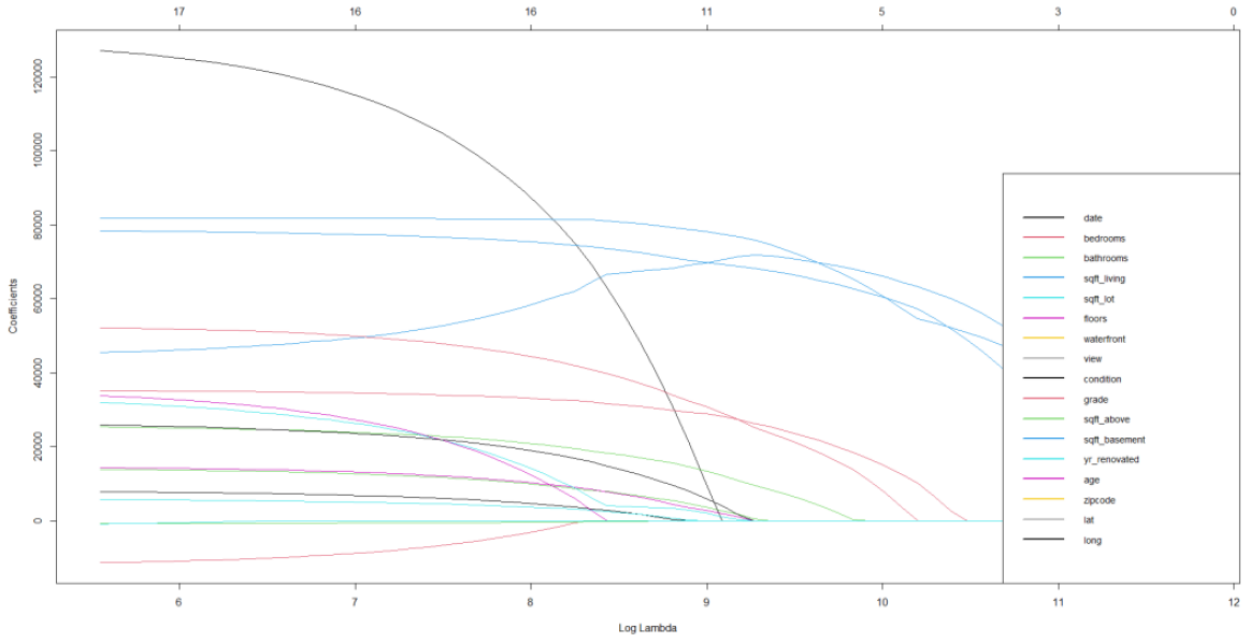


Рисунок 3.37 – Коефіцієнти лассо регресії при збільшенні значення λ

Аналогічно до рідж-регресії, для визначення оптимального значення λ ми виконуємо процедуру крос-валідації. Для цього ми застосовуємо той самий підхід, що і в рідж-регресії, змінюючи значення альфа на 1. Ми спостерігаємо, що MSE може бути мінімізовано приблизно в діапазоні $0 \leq \log(\lambda) \leq 8$. Це не тільки знижує значення MSE, але також зменшує кількість змінних з 17 до 16, як видно на діаграмі 3.38.

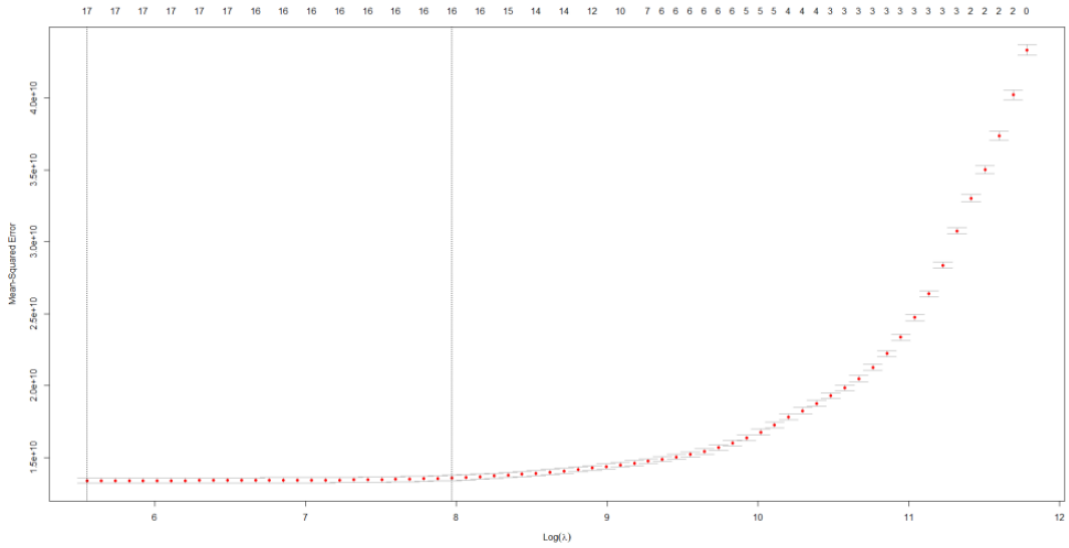


Рисунок 3.38 – Результат перехресної перевірки для визначення оптимального значення λ

Перша та друга вертикальні пунктирні лінії представляють значення λ, при яких відбувається мінімізація середньоквадратичної помилки (MSE), відповідно до найменшого та найбільшого значень λ в межах однієї стандартної помилки мінімального MSE. Були отримані наступні значення:

– Мінімальне MSE = 13,403,564,846, λ = 257.3681
– Стандартна похибка мінімального MSE = 13,678,332,513, λ = 3,482.313

На підставі цих значень та даних, наведених на рисунку 3.39, можна зробити висновок, що використання λ, розташованого на відстані однієї стандартної похибки, є більш підходящим, оскільки λ, що відповідає мінімальному MSE, майже не впливає на змінні.

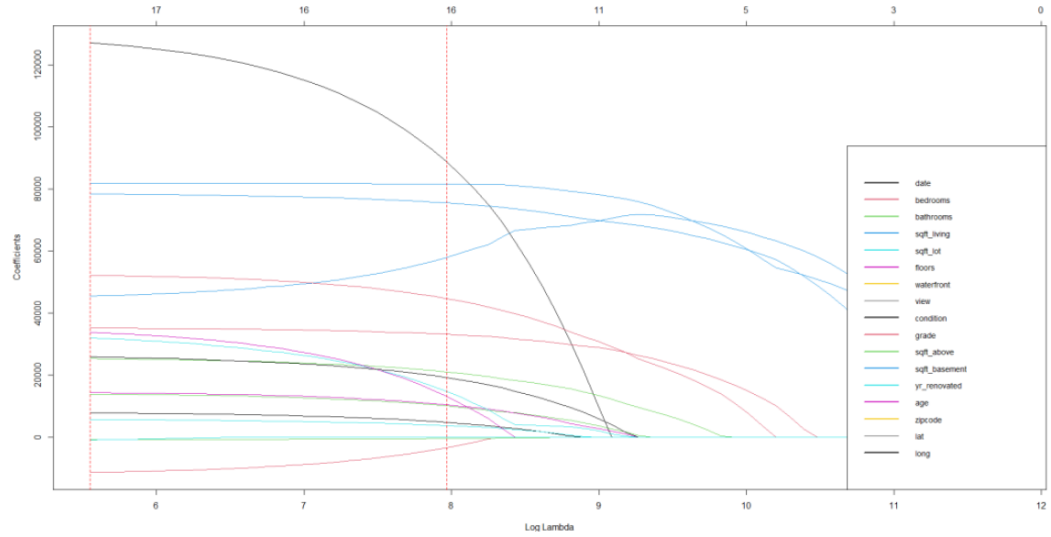


Рисунок 3.39 – Оптимальне значення λ

Наступним кроком було побудувати модель ласо з використанням параметра альфа, який дорівнює одній стандартній похибці, як показано на рис. 3.40. З цього можна зробити висновок, що 68,6% дисперсії залежної змінної можна пояснити незалежними змінними.

```
Call:  glmnet(x = x_train, y = y_train, alpha = 1, lambda = lasso$lambda.1se,      standardize = TRUE)

  Df %Dev Lambda
1 16 68.6   3482
```

Рисунок 3.40 – Результат лассо регресії

Останнім етапом є перевірка ефективності на тестовому наборі даних. Після обчислень можна визначити середньоквадратичну помилку (MSE) та коефіцієнт детермінації (R^2) для моделі. Значення отримані такі: MSE = 13500403033, $R^2 = 0.69$.

Останньою моделлю регуляризації, яка була використана, є еластична сітка. Ми реалізуємо еластичну сітку так само, як моделі рідж та ласо, з використанням параметра альфа. Значення альфа від 0 до 1 визначає тип регуляризації: при альфа = 0,5 використовується комбінація обох штрафів.

На рисунку 3.41 можна побачити, що при log(λ) = 6 всі 16 змінних використовуються в моделі, при log(λ) = 8 використовується 15 змінних, а при log(λ) = 11 лише 4 змінних зберігаються.

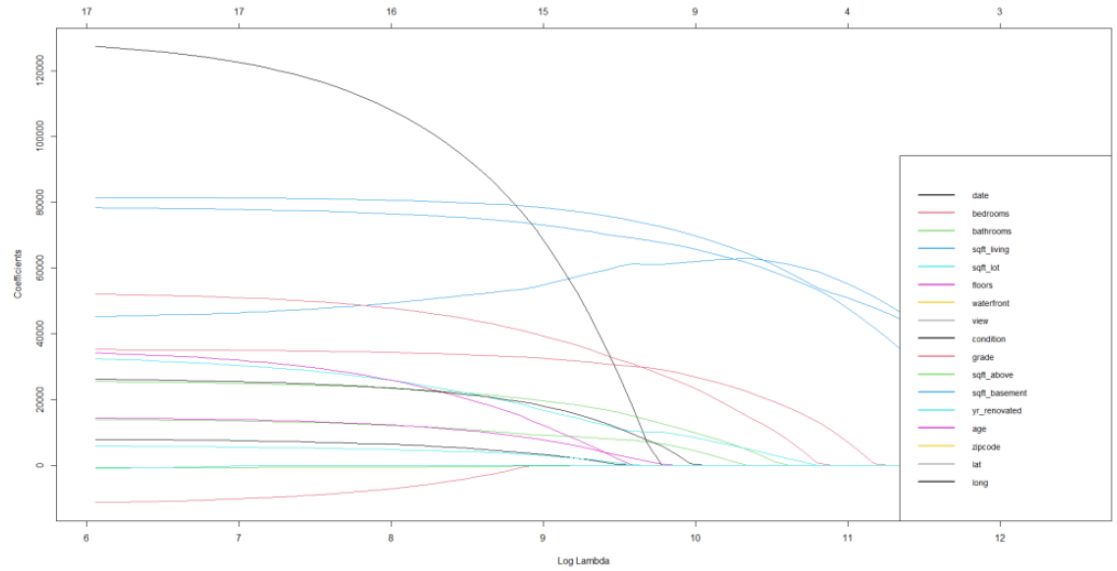


Рисунок 3.41 – Коефіцієнти еластичної сітки при збільшенні значення λ

Так само як у випадку рідж та ласо регресій, нам необхідно виконати крос-валідацію, щоб визначити оптимальне значення λ. Для CV ми використовуємо той самий підхід, але змінюємо наш параметр α на 0.5. З рисунку 3.42 видно, що ми можемо досягти найменшого середньоквадратичного помилки (MSE) приблизно в діапазоні $6 \leq \log(\lambda) \leq 9$. Це не тільки мінімізує наше MSE, але також зменшує кількість змінних з 17 до 16.

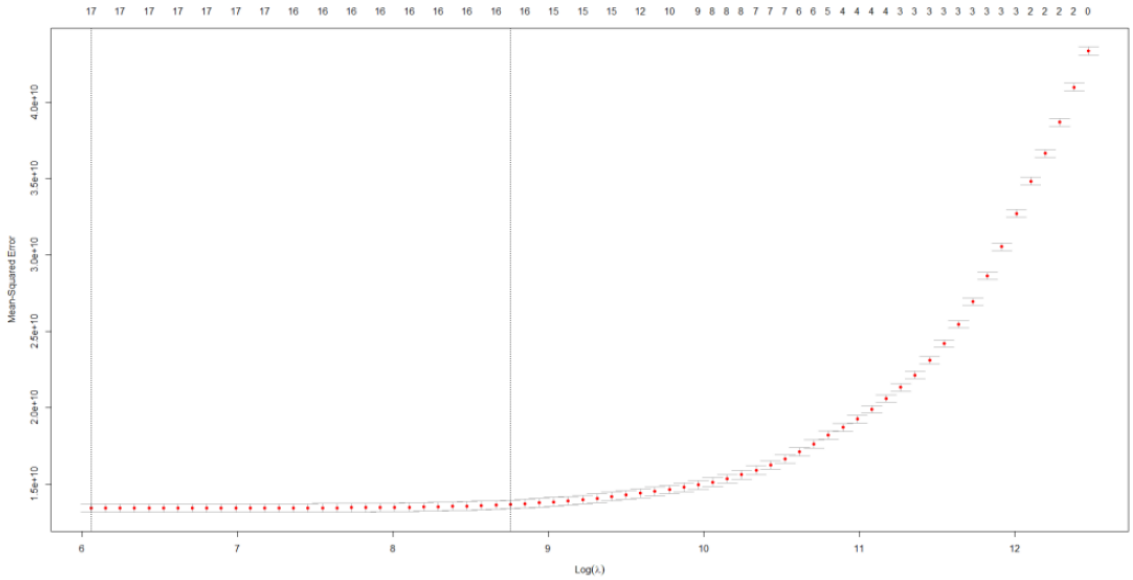


Рисунок 3.42 – Результат перехресної перевірки для визначення оптимального значення λ

Перша і друга вертикальні пунктирні лінії на рисунку представляють мінімальне значення MSE і найбільше значення λ, відповідно, що знаходяться в межах однієї стандартної похибки мінімального значення MSE. Були отримані такі значення:

– Мінімальне MSE = 13404487261, λ = 427,343
– Стандартна похибка мінімального MSE = 13574829564, λ = 5268,484

Таким чином, з рисунку 3.43 можна сказати, що доцільніше використовувати λ, яке відповідає одній стандартній похибці, оскільки λ мінімального MSE практично не впливає на змінні.

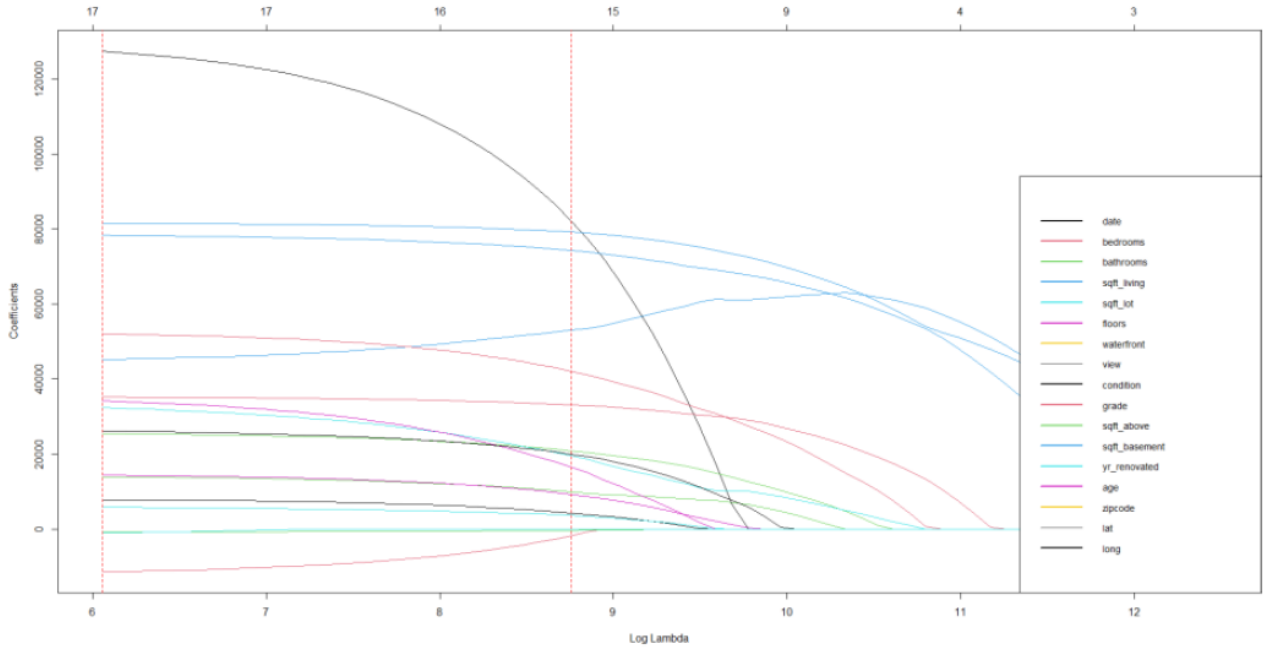


Рисунок 3.43 – Оптимальне значення λ

Далі було побудовано модель еластичної сітки з використанням параметра альфа, що дорівнює 0,5, та лямбди, яка відповідає одній стандартній похибці, як показано на рис. 3.44. З цього можна зробити висновок, що 68,83% дисперсії залежної змінної можна пояснити за допомогою незалежних змінних.

```
Call:  glmnet(x = x_train, y = y_train, alpha = 0.5, lambda = elastic_net$lambda.1se,      standardize = TRUE)

  Df  %Dev Lambda
1 16 68.83   5268
```

Рисунок 3.44 – Результат еластичної сітки

Останнім етапом є оцінка ефективності моделі за допомогою тестового набору даних. Після обчислень можна отримати середньоквадратичну помилку (MSE) і коефіцієнт детермінації (R квадрат). Значення, які ми отримали, такі: MSE = 13424321775, R квадрат = 0,692. Далі ми порівняємо всі побудовані моделі і представимо результат на рисунку 3.45.

Зробимо висновок, що модель еластичної сітки має найбільше значення R квадрат та найменше значення MSE серед усіх розглянутих моделей з регуляризацією. Тому значення цієї моделі будуть порівнюватися з результатами інших моделей на заключному етапі.

	R-squared	Mse
ridge cross-validated	0.6910389	13512244229
lasso cross_validated	0.6898402	13500403033
elastic net cross_validated	0.6915537	13424321775

Рисунок 3.45 – Оцінки ефективності моделей регуляризації

Тому ми використовуємо модель прогнозу на основі даних моделі еластичної сітки, щоб оцінити точність. Отримані результати такі:

– Коренева середня квадратична помилка (RMSE) = 115863,4
– Середня абсолютна помилка (MAE) = 87478,46
– Відсоток середньої абсолютної відносної помилки (MAPE) = 0,203577
– Симетричний відсоток середньої абсолютної відносної помилки (SMAPE) = 0,1926165
– Масштабована абсолютна помилка (MASE) = 0,3753061

Середньоквадратична помилка (RMSE) вказує, що прогнозовані значення в середньому відхиляються на 115863,4 доларів від експериментальних значень. Показник середньої відсоткової абсолютної помилки (MAPE) показує, що прогнозовані значення в середньому відхиляються на 20,4% від експериментальних значень.

3.4 Моделювання цін на нерухомість методом Random Forest

Наступна модель, яку було побудовано, - модель випадкового лісу. Початково на тренувальному наборі даних було створено базову модель випадкового лісу, яка за замовчуванням будує 500 дерев та випадковим чином вибирає 1/3 всіх незалежних змінних при кожному розбитті.

За допомогою рис. 3.46 можна зробити висновок, що кількість змінних, випробуваних при кожному розбитті, становить 5; середньоквадратична помилка (MSE) дорівнює 6215759789 (або корінь середньоквадратичної помилки - RMSE = 78840), і 85,68% дисперсії залежної змінної можна пояснити незалежними змінними.

```
Call:
 randomForest(formula = price ~ ., data = train, proximity = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 6215759789
                    % Var explained: 85.68
```

Рисунок 3.46 – Результат моделювання випадкового лісу

На рисунку 3.47 спостерігається стабілізація нашого коефіцієнта помилок приблизно після використання 100 дерев, але його подальше зменшення відбувається повільніше, наближаючись до 300 дерев. Графік коефіцієнта помилок базується на помилках вибірки, тому ми можемо визначити кількість дерев, при якій досягається найнижчий коефіцієнт помилок.

За результатами дослідження було встановлено, що 456 дерев забезпечують середню похибку ціни продажу житла в розмірі 78788,97 доларів.

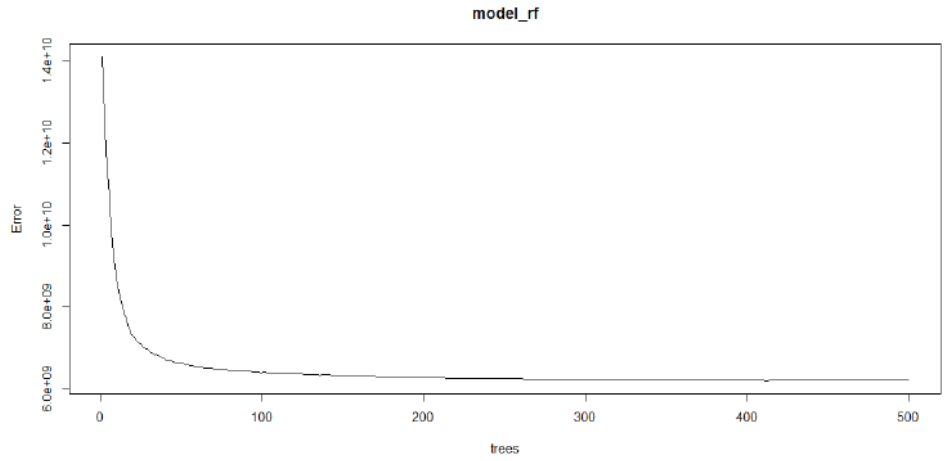


Рисунок 3.47 – Коефіцієнт помилок випадкового лісу

Після проведення розрахунків наступним кроком є оцінка ефективності на тестовому наборі даних. Це дозволить визначити наступні показники ефективності:

– Коефіцієнт детермінації (R^2) = 0,863
– Корінь середньоквадратичної помилки (RMSE) = 77621,1
– Середня абсолютна помилка (MAE) = 53353,94
– Відносна абсолютна помилка (MAPE) = 0,1231881
– Симетрична відносна абсолютна помилка (SMAPE) = 0,1163658
– Стандартизована абсолютна помилка (MASE) = 0,2289027

Проте, ми можемо продовжити пошук вдосконалень, налаштувавши нашу модель. Гіперпараметр - коригуючий параметр, що використовується для керування процесом навчання. Значення гіперпараметрів потрібно встановити перед початком навчального процесу.

Якщо ми здійснюємо пошук за декількома гіперпараметрами, то такий процес називається "пошук по сітці", де вичерпно перевіряються всі можливі комбінації значень гіперпараметрів для заданого набору параметрів і значень.

На рисунку 3.48 показана сітка, що складається з наступних гіперпараметрів:

- ntrees: кількість дерев, що створюються для регресії.

- mtries: на кожній ітерації випадковим чином вибирається підмножина ознак з навчальних даних для визначення оптимального розподілу цієї підмножини.

- max_depth: максимальна глибина дерева, тобто найбільша кількість шляхів від кореня до листка.

- min_rows: мінімальна кількість спостережень для кожного термінального вузла.

- nbins: кількість груп, на які можна розділити вихідні дані.

- sample_rate: відсоток вибірки рядків для кожного дерева.

```
hyper_grid.h2o <- list(
  ntrees      = seq(50, 500, by = 50),
  mtries      = seq(5, 15, by = 10),
  max_depth   = seq(20, 40, by = 5),
  min_rows    = seq(1, 5, by = 1),
  nbins       = seq(10, 30, by = 5),
  sample_rate = c(.55, .632, .75)
)
```

Рисунок 3.48 – Параметри моделі для пошуку по сітці

Після оцінки 35 моделей, була вибрана найкраща комбінація гіперпараметрів, при яких значення MSE є найменшим. Згідно з результатами, представленими на рисунку 3.49, оптимальною моделлю є та, яка має такі характеристики: 450 дерев, 15 вибраних змінних на кожному рівні, максимальна глибина 20, вибір 1 спостереження в кожному термінальному вузлі, розділення даних на 30 груп та частота дискретизації рядка 0,56.

З використанням отриманих параметрів можна побудувати модель на основі навчального набору даних.

```
best_model <- h2o.randomForest(
  x = x,
  y = y,
  training_frame = train.h2o,
  ntrees      = 450,
  mtries      = 15,
  max_depth   = 20,
  min_rows    = 1,
  nbins       = 30,
  sample_rate = .55)
```

Рисунок 3.49 – Результати пошуку по сітці

Після проведення експерименту з випадковим лісом виникає природне питання про те, які змінні найбільше впливають на прогнозовану силу моделі. Змінні, які мають велике значення, визначають результат і мають суттєвий вплив на його значення. Зворотно, змінні з низькою вагомістю можуть бути виключені з моделі, що спрощує її налаштування та прогнозування.

Співвідношення важливості змінних на рисунку 3.50 вказують на те, які змінні мають найбільший вплив. Площа житла ("sqft_living") та широта ("lat") є найбільш вагомими, тоді як "sqft_basement" (площа підвалу), "waterfront" (вихід до набережної), "yr_renovated" (рік останнього ремонту) є найменш вагомими.

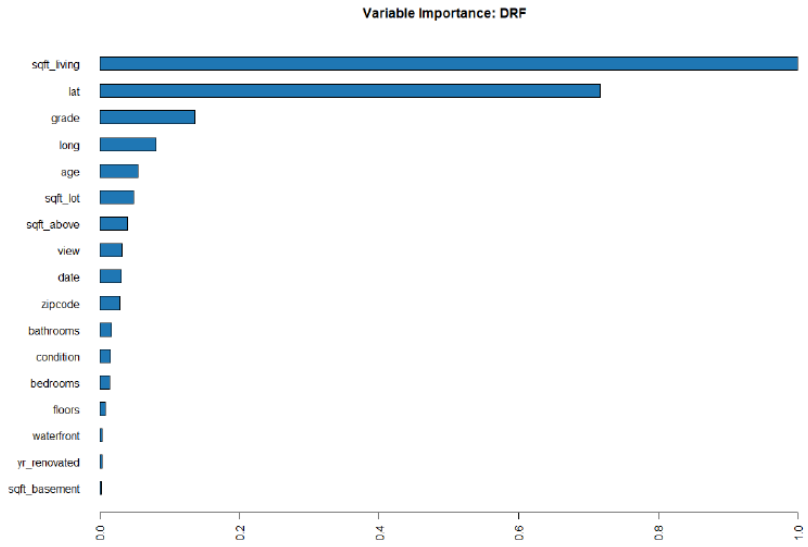


Рисунок 3.50 – Відносна важливість змінних

Після аналізу продуктивності на експериментальному датасеті, були отримані такі показники:

– Коефіцієнт детермінації (R^2) дорівнює 0,864, що вказує на те, що приблизно 86% дисперсії залежної змінної можна пояснити незалежними змінними.

– Середньоквадратична помилка (RMSE) складає 76899,5, що означає, що прогнозовані значення в середньому відхиляються від експериментальних на 76899,5 доларів.

– Відносна абсолютна помилка (MAPE) становить 0,11122, що свідчить про те, що прогнозовані значення в середньому відхиляються від експериментальних на 11%.

Останнім етапом є порівняння двох моделей за допомогою показників R^2 та RMSE, що можна побачити на рис. 3.51. Зазначені показники свідчать про те, що модель з налаштованими параметрами є кращою.

	R-squared	RMSE
Default RF	0.8626981	77621.1
RF with Tuning HP	0.8639315	76899.5

Рисунок 3.51 – Оцінки ефективності моделей випадкового лісу

3.5 Моделювання цін на нерухомість методом XGBoost

На завершення, модель, побудована за допомогою алгоритму машинного навчання XGBoost, є аналогічною до моделі випадкового лісу в тому, що параметри відіграють важливу роль у її конструюванні. Рекомендується спочатку спробувати побудувати модель XGBoost з використанням перехресної перевірки та значеннями за замовчуванням для таких параметрів:

eta - контролює швидкість навчання та робить модель більш надійною шляхом зменшення ваги на кожному кроці. Значення за замовчуванням: 0,3.

max_depth - глибина дерева. Значення за замовчуванням: 6.

min_child_weight - мінімальна кількість спостережень, необхідних у кожному термінальному вузлі. Значення за замовчуванням: 1.

subsample - відсоток навчальних даних для вибірки для кожного дерева. Значення за замовчуванням: 100%.

За такими параметрами була побудована базова модель XGBoost з 1000 деревами, яка була оцінена за допомогою 10-кратної перехресної перевірки. Після цього ми маємо можливість оцінити цю модель для визначення мінімального RMSE (середньоквадратична помилка) та оптимальної кількості дерев як для навчальних даних, так і для перехресно-перевіреної помилки.

З рисунку 3.52 видно, що помилка навчання продовжує зменшуватися до 1000 дерев, де RMSE майже досягає значення 4492,162 доларів. Однак перехресно-перевірена помилка досягає мінімального RMSE у 76240,48 доларів США лише з використанням 134 дерев.

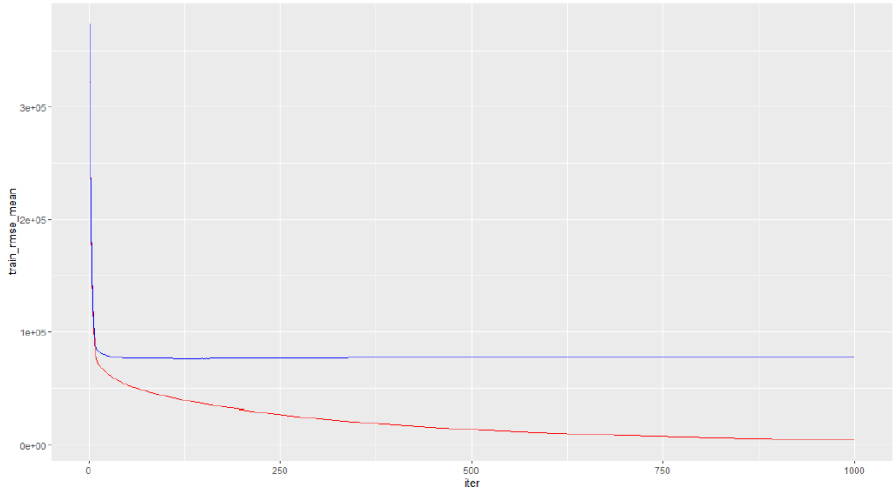


Рисунок 3.52 – Мінімальне значення RMSE та оптимальна кількість дерев як для навчальних даних, так і для перехресно перевіреної помилки

Зміни в значеннях гіперпараметрів можуть бути використані для налаштування моделі. Використання пошуку по сітці допоможе встановити оптимальні значення параметрів, що забезпечують найкращу оцінку RMSE. Для цього було запущено 10-кратну перехресну перевірку з використанням 5000 дерев та пошуком параметрів, які вказані на рис. 3.53.

У додаток до вказаних параметрів, в наступній моделі будуть використовуватися також параметри: gamma, який сприяє обрізці дерев, lambda, який відповідає за регуляризацію, а також colsample_bytrees, який вказує на відсоток стовпців, які використовуються для кожного дерева.

```
hyper_grid <- expand.grid(
  eta = c(.01, .05, .1, .3),
  gama = c(0,.25,1),
  lambda = c(0,1,10),
  max_depth = c(1, 3, 5, 7),
  min_child_weight = c(1, 3, 5, 7),
  subsample = c(.65, .8, 1),
  colsample_bytree = c(.75, .9, 1),
  optimal_trees = 0,
  min_RMSE = 0
)
```

Рисунок 3.53 – Параметри моделі

Після виконання пошуку було встановлено, що використання параметрів, зображених на рисунку 3.54, привело до зниження значення RMSE до 70812,76 доларів. Після цього модель можна застосувати до тренувальних даних, використовуючи 2018 дерев та вибрані параметри.

```
params <- list(
  eta = 0.01,
  gama = 0.25,
  lambda = 5,
  max_depth = 7,
  min_child_weight = 1,
  subsample = 0.65,
  colsample_bytree = 0.75
)
# train final model

xgb.fit.final <- xgboost(
  params = params,
  data = x_train,
  label = y_train,
  nrounds = 2018,
  objective = "reg:linear",
  verbose = 0
)
```

Рисунок 3.54 – Введення параметрів для пошуку по сітці

На рисунку 3.55 можна спостерігати відносну вагомість змінних у моделі. Найбільш значущими змінними є широта (lat), житлова площа (sqft_living) та індекс завідповідності будівельному та дизайнерському рівням (grade), тоді як найменш значущими є вихід на набережну (waterfront), рік останнього ремонту (yr_renovated) та кількість поверхів (floors).

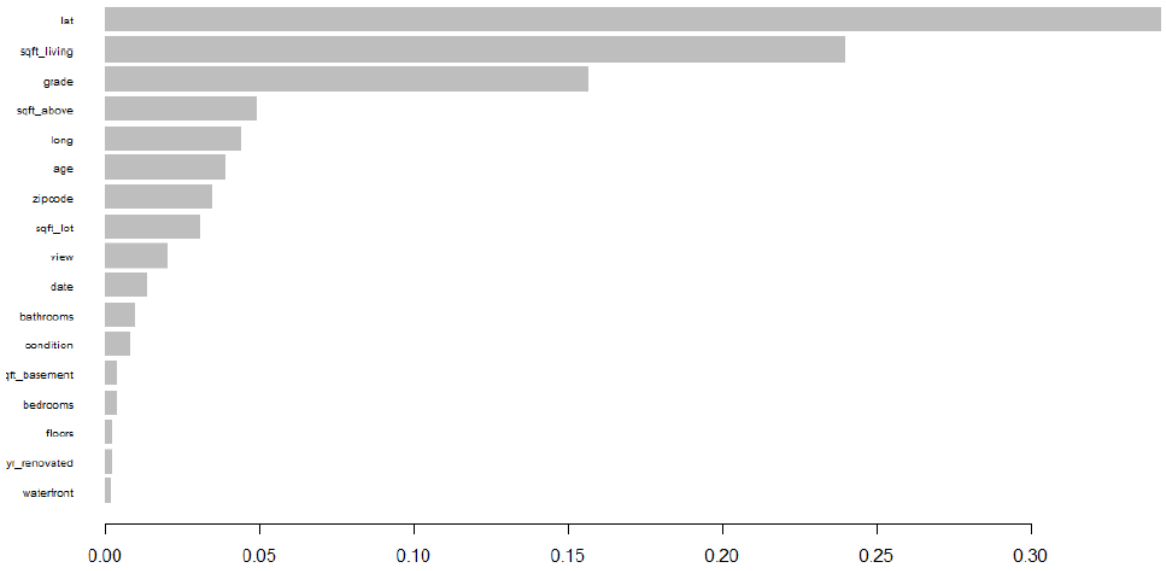


Рисунок 3.55 – Відносна важливість змінних

В порівнянні з відносною важливістю змінних у моделі випадкового лісу, можна помітити, що значення важливості змінних різняться. Це є очікуваним, оскільки метод пристосування дерев у кожній моделі може варіюватися. Однак, незалежно від моделі, змінні, що відносяться до широти та площі будинку, схоже, мають значний вплив на ціну житла. Цей факт підтверджує, що ці дві змінні є основними факторами, що визначають вартість нерухомості. Після обчислень можна отримати такі показники перевірки ефективності на тестовому датасеті:

- Коефіцієнт детермінації (R-квадрат) = 0,893
- Корень середньої квадратичної помилки (RMSE) = 67669,15
- Середня абсолютна помилка (MAE) = 47063,49
- Відсоткова абсолютна помилка (MAPE) = 0,1034903
- Симетрична відсоткова абсолютна помилка (SMAPE) = 0,1048065
- Середня абсолютна помилка рівня (MASE) = 0,2188252

Порівнявши значення всіх побудованих моделей за алгоритмом XGBoost, найкращі показники демонструє модель, чиї параметри були знайдені за допомогою пошуку по сітці.

3.6 Порівняння отриманих результатів та пояснення

Під час проведених досліджень було обрано найкращу модель для порівняння серед кожного застосованого алгоритму. За результатами, які можна побачити на рис. 3.56, можна зробити такі висновки:

1. Використання повної лінійної регресійної моделі дозволяє пояснити близько 83,4% дисперсії залежної змінної за допомогою незалежних змінних. В середньому, отримані значення від прогнозу відхиляються від значень, отриманих в експерименті, на 85 054,44 доларів або 15%.

2. Використання моделі еластичної сітки дозволяє пояснити близько 69,1% дисперсії залежної змінної за допомогою незалежних змінних. В середньому, отримані значення від прогнозу відхиляються від значень, отриманих в експерименті, на 115 863,40 доларів або 20%.

3. Використання алгоритму випадкового лісу дозволяє пояснити близько 86,4% дисперсії залежної змінної за допомогою незалежних змінних. В середньому, отримані значення від прогнозу відхиляються від значень, отриманих в експерименті, на 76 961,49 доларів або 14%.

4. Використання алгоритму XGBoost дозволяє пояснити близько 89,3% дисперсії залежної змінної за допомогою незалежних змінних. В середньому, отримані значення від прогнозу відхиляються від значень, отриманих в експерименті, на 67 669,15 доларів або 10%.

	R-squared	RMSE	MAPE
Full Linear model	0.834	85054.44	0.15
Elastic Net	0.691	115863.40	0.20
Random Forest	0.864	76961.49	0.14
XGBoost	0.893	67669.15	0.10

Рисунок 3.56 – Оцінка показників ефективності створених моделей

Таким чином, найкращим алгоритмом машинного навчання для прогнозування ціни на цьому наборі даних є XGBoost, оскільки він демонструє лише 10% відхилення прогнозованих значень від реальних, що є дуже добрим показником.

Лінійна регресія може бути корисним інструментом для вивчення взаємозв'язків між змінними, але для більшості практичних випадків не рекомендується через свою спрощену природу. Цей метод припускає лінійну залежність між змінними, що не завжди відповідає реальним проблемам. Навіть у моєму прикладі половина вхідних параметрів була категорійною та не мала прямої лінійної залежності від ціни.

Крім того, лінійна регресія передбачає незалежність даних, що не завжди є реалістичним. Навіть при спробі уникнути мультиколінеарності, це може бути непростим завданням, і оцінки ваги параметрів можуть бути невірними, оскільки залежать від взаємозв'язку з іншими незалежними змінними.

Алгоритм машинного навчання XGBoost, який використовує як лінійні так і нелінійні відносини, працює на основі ансамблевого навчання. Це дозволяє поєднати сильні сторони різних моделей для отримання кращих результатів. Я вважаю, що основною перевагою цього алгоритму порівняно з іншими є можливість використовувати різноманітні гіперпараметри, наприклад, для регуляризації та управління складністю дерев.

Оскільки XGBoost використовує дерева прийняття рішень, він має властивість імунітету до мультиколінеарності. Наприклад, якщо деякі ознаки в даних сильно корельовані, алгоритм автоматично обере лише одну з них на кожному розбитті.

Щоб порівняти ці два методи і визначити, який з них краще передбачає ціну, я випадковим чином обрав будинок з характеристиками, які можна побачити на рисунку 3.57:

id	11209
price	427 800
date	14.11.2018
bedrooms	3
bathrooms	1,75
sqft_living	1340
sqft_lot	13241
floors	1
waterfront	0
view	2
condition	3
grade	7
sqft_above	1340
sqft_basement	0
yr_built	1985
yr_renovated	0
age	36
repair_age	0
zipcode	98034
lat	47,727
long	-122,21

Рисунок 3.57 – Характеристики випадково обраного будинку

Після внесення вхідних даних до обох моделей, таблиця 3.5 відображає наступні висновки:

Таблиця 3.5 – Результати прогнозування ціни за лінійної моделі та моделі XGBoost

	Реальна ціна	Прогнозована ціна	Різниця
Лінійна модель	427800 $	451232,6 $	-23432,6
XGBoost	427800 $	425377,8 $	2422,2

За лінійною моделлю виявлено переоцінку ціни будинку на 23432,6 долара, тоді як за XGBoost ціна була недооцінена на 2422,2 долара. Ці результати свідчать про те, що модель XGBoost краще відображає дані і має вищу ймовірність прогнозування реальної ціни. Проте, в обох випадках ситуація не є ідеальною, оскільки як завищені, так і занижені ціни є несприятливими для продажу будинків.

Якби ціна на будинок визначалась за лінійною моделлю, це означало б, що ціна будинку надмірно завищена. Такий сценарій має негативні наслідки, оскільки може створити численні перешкоди, серед яких:

- Зниження попиту покупців: якщо вартість будинку значно перевищує ринкову ціну, існує ризик відлякати покупців і суттєво зменшити попит.

- Продовження тривалого перебування на ринку: чим довше будинок знаходиться на ринку, тим менш привабливим він стає для потенційних покупців. Якщо будинок довго продається, це може викликати підозри інтересованих осіб, що з будинком щось не так. Крім того, тривале перебування на ринку підвищує ризик негативних змін на ринку. У разі зниження ринкових цін може знадобитися значне зниження ціни продажу, що призводить до більших втрат.

- Зменшення ціни: кожне нове зниження ціни на житло викликає підозри серед покупців. Якщо ціна постійно падає, це може викликати запитання про якість самого будинку.

- Витрати на утримання: розміщення оголошень про продаж може бути дорогою процедурою. Багато продавців наймають професійні прибирання для своїх будинків, і чим довше будинок знаходиться на ринку, тим більше грошей потрібно витратити на маркетинг, що збільшує загальні витрати [73].

Таким чином, якщо ціна була б вищою на 23432,6 долара, існувала б велика ймовірність того, що будинок не було б куплено на тривалий період, а продавець, крім втрати прибутку, мусів би зіткнутися з додатковими витратами або навіть знизити ціну нижче ринкової, що привело б до ще більших втрат.

З іншого боку, якщо ціна була встановлена за допомогою алгоритму XGBoost, будинок міг би мати незначно занижену ціну. Хоча цей сценарій не є найкращим, багато експертів рекомендують не боятися встановлювати нижчу ціну, оскільки теоретично це може збільшити кількість пропозицій і привести ціну до фактичної ринкової вартості будинку.

Таким чином, якби ціна була на 2422,2 долара меншою, продавець заробив би трохи менше, але така різниця не є критичною, і швидкий продаж уберіг би від додаткових витрат. З економічної точки зору, прогнозування ціни з використанням алгоритму машинного навчання XGBoost є більш точним, що дозволяє продавцю встановити правильну ціну на будинок і менше ризикувати втратою грошей.

РОЗДІЛ 4. ПРАКТИЧНА ЦІННІСТЬ ТА МОЖЛИВОСТІ ВИКОРИСТАННЯ

Використання методів машинного навчання для прогнозування цін на нерухомість є потрібним з кількох причин:

1. Точність прогнозування: Машинне навчання може використовувати складні алгоритми та моделі, які дозволяють точніше прогнозувати ціни на нерухомість. Враховуючи широкий спектр факторів, таких як розмір властивості, розташування, історичні дані та інші характеристики, модель може здійснювати прогнози з високою точністю.

2. Ефективне використання даних: Велика кількість даних є доступною для аналізу в нерухомість, включаючи історичні дані про продажі, оренду, характеристики властивості, економічні показники та багато іншого. Машинне навчання може ефективно обробляти ці дані, знаходити кореляції та залежності, що допомагає покращити якість прогнозів.

3. Виявлення складних зв'язків: Ринок нерухомості є дуже складним і має багато взаємозв'язків між факторами, які впливають на ціни. Машинне навчання може виявляти складні залежності та неочевидні зв'язки між різними факторами, що допомагає отримати більш точні та об'єктивні прогнози.

4. Швидкість та ефективність: Машинному навчанню не потрібно проводити довготривалі аналітичні обчислення або ручні розрахунки. Воно може швидко обробити велику кількість даних та забезпечити швидкі та ефективні прогнози цін на нерухомість.

5. Прийняття обґрунтованих рішень: Прогнозування цін на нерухомість за допомогою машинного навчання надає корисну інформацію та аналіз, які допомагають робити обґрунтовані рішення щодо інвестицій у нерухомість, продажу або покупки власної нерухомості.

Багато компаній використовують методи машинного навчання для обробки даних ринку нерухомості. Деякі з них включають:

1. Zillow: Zillow використовує машинне навчання для прогнозування цін на нерухомість, виявлення трендів ринку та покращення точності оцінок нерухомості [74].

2. Redfin: Redfin також використовує методи машинного навчання для аналізу ринкових даних та прогнозування цін на нерухомість. Вони використовують моделі глибокого навчання для виявлення закономірностей та трендів на ринку [75].

3. Trulia: Trulia використовує машинне навчання для аналізу ринкових даних, прогнозування цін на нерухомість та рекомендацій для користувачів. Вони також використовують нейронні мережі для покращення точності своїх прогнозів [76].

4. Compass: Compass використовує методи машинного навчання для аналізу даних ринку нерухомості та прогнозування цін. Вони використовують алгоритми машинного навчання, щоб зрозуміти зв'язки між різними факторами та цінами на нерухомість [77].

5. Realtor.com: Realtor.com використовує машинне навчання для аналізу ринкових даних, класифікації нерухомості та рекомендацій для користувачів. Вони використовують алгоритми навчання з підкріпленням для покращення своїх прогнозів [78].

Це лише кілька прикладів компаній, які використовують методи машинного навчання для обробки даних ринку нерухомості. Список не є вичерпним, оскільки багато компаній у цій галузі використовують аналогічні методи для аналізу та прогнозування ринку нерухомості.

Для обраного підприємства, що являється базою практики, було розроблено план по впровадженню технології аналізу та прогнозування цін на нерухомість.

Перший крок - це збір відповідних даних про ринок нерухомості. Це можуть бути дані про ціни продажу та оренди, характеристики власності, географічні дані, тенденції ринку, демографічні дані тощо. Ці дані можна отримати з різних джерел, включаючи місцеві бази даних, інтернет-ресурси, звіти ринку, джерела нерухомості тощо.

Існує кілька інструментів і джерел, які можна використовувати для збору даних про нерухомість. Ось деякі з них:

1. Місцеві агентства нерухомості: Звернення до місцевих агентств нерухомості може бути одним з найпростіших способів отримати дані про ціни на нерухомість, характеристики власності та ринкові тенденції. Агентства надають інформацію про наявні об'єкти, їх вартість, умови продажу або оренди.

2. Онлайн-ресурси та веб-сайти: Інтернет-ресурси, такі як сайти оголошень про нерухомість, платформи для продажу та оренди, агрегатори даних про нерухомість, можуть бути цінним джерелом інформації. Такі ресурси зазвичай надають дані про ціни, характеристики об'єктів, фотографії та інші відомості.

3. Громадські реєстри: Деякі країни мають громадські реєстри нерухомості, які зберігають інформацію про власності, права власності, транзакції та інші дані про нерухомість. Ці реєстри можуть бути доступні онлайн або в офісах реєстрації.

4. Державні органи та організації: Деякі державні органи та організації, такі як національні статистичні агентства, регулятори нерухомості, місцеві управління та планування, можуть надавати дані про ринок нерухомості, статистику продажів, будівництва, дозволи на будівництво та інші важливі відомості.

5. Експертні думки та консультанти: Звернення до експертів у сфері нерухомості, таких як оцінювачі, риночники, аналітики або консультанти, може забезпечити цінну інформацію про ринок та тенденції. Вони можуть мати

доступ до недоступних загальному публіці даних та мати глибокі знання про ринок нерухомості.

Ці інструменти можуть бути використані окремо або в поєднанні для збору даних про нерухомість. Важливо забезпечити достовірність та актуальність даних, а також врахувати контекст і особливості ринку нерухомості при зборі та аналізі інформації.

Після збору даних необхідно провести їх аналіз і підготовку. Це включає очищення даних від відсутніх значень, виявлення та виправлення аномалій, агрегацію даних за певними параметрами, які можуть вплинути на ціни нерухомості (наприклад, площа, кількість кімнат, розташування тощо).

Після підготовки даних необхідно використати метод XGBoost як найбільш доцільний, виходячи з результатів дослідження. Наступним кроком є побудова прогнозної моделі на основі обраного методу. Це включає визначення залежних та незалежних змінних, розбиття даних на тренувальний та тестовий набори, побудову моделі за допомогою вибраного методу та підгонку моделі до тренувальних даних.

Після побудови моделі її необхідно протестувати на тестовому наборі даних, щоб оцінити її точність та ефективність. Якщо модель дає задовільні результати, її можна використовувати для прогнозування цін на нерухомість. Прогнозні моделі потребують постійного моніторингу та оновлення відповідно до змін на ринку нерухомості. Ринкові тенденції, нові дані та інші фактори можуть вплинути на ціни нерухомості, тому моделі слід періодично оновлювати для забезпечення точності і актуальності прогнозів.

Для відстеження змін на ринку нерухомості існує кілька методів і інструментів, які можна використовувати. Ось декілька способів:

1. Аналіз ринкових звітів: Багато компаній, брокерських агентств, дослідницьких організацій та нерухомісних консалтингових фірм публікують регулярні звіти про стан ринку нерухомості. Ці звіти надають інформацію про

ціни, тенденції, обсяги продажів, попит та інші ключові показники ринку. Відстеження таких звітів допомагає отримати уявлення про загальну ситуацію на ринку.

2. Моніторинг оголошень про нерухомість: Перегляд онлайн-оголошень про нерухомість на платформах продажу та оренди може дати уявлення про зміни в цінах, доступність об'єктів, пропозиції та попит на ринку. Відстежування оголошень може допомогти спостерігати, які типи нерухомості є популярними та як їхні ціни змінюються з часом.

3. Взаємодія з професіоналами: Спілкування з ріелторами, агентами нерухомості, оцінювачами та іншими фахівцями ринку може дати цінну інформацію про зміни на ринку. Вони можуть мати першу руку на дані про нові проекти, тенденції попиту та інші фактори, що впливають на ринок нерухомості.

4. Використання даних з громадських реєстрів: У деяких країнах існують громадські реєстри нерухомості, які містять інформацію про продажі, транзакції та вартість нерухомості. Можливість доступу до таких даних дозволяє відстежувати зміни в цінах та тенденціях на ринку.

5. Аналіз економічних показників: Економічні показники, такі як ставки відсотка, індекси споживчої ціни, зміни зайнятості та економічний ріст, можуть впливати на ринок нерухомості. Відстеження таких показників дозволяє розуміти, які фактори впливають на ринок та як це може відобразитися на цінах нерухомості.

Ці методи можуть бути використані окремо або в поєднанні, залежно від доступності даних та вимог спостерігача на ринку нерухомості.

Таким чином, використання результатів дослідження та спостереження за актуальними змінами ринку нерухомості дають можливість підвищити ефективність роботи підприємства, в даному випадку бази практики. Витрати часу та ресурсів на вивчення проблеми та імплементацію рішення на

підприємстві, в майбутньому дозволять зекономити час на обробку інформації та аналіз, збільшать показники прибутку і дадуть простір для експериментів.

ВИСНОВКИ

В роботі було встановлено, що нерухомість означає власність, яка складається з землі та будівель, що включають земельну ділянку та нерухомі об'єкти. Ринок нерухомості включає житлову нерухомість, яка продається та купується безпосередньо між покупцями та продавцями або через посередників. Він є важливою складовою соціально-економічної системи і залучає широке коло зацікавлених осіб, починаючи з покупців та продавців, і закінчуючи приватними та державними установами.

Ціна на житлову нерухомість, подібно до будь-якого активу, залежить від закону попиту та пропозиції. Коли ціни на будинки знижуються, все більше людей розглядають можливість придбання власного житла, що призводить до збільшення попиту. Зі зростанням ціни на нерухомість збільшується кількість доступних на ринку пропозицій. Зміни на ринку житла завжди мають важливе значення, оскільки вони взаємозв'язані з цінами та споживчими витратами. Спостереження за ринком житла допомагають оцінити загальний попит на товари та послуги. Крім того, купівля та продаж будинків мають безпосередній вплив на економіку, оскільки інвестиції в житло є частиною загального обсягу виробництва в економіці.

Дослідження сучасних тенденцій розвитку ринку нерухомості вказало на те, що в найближчому майбутньому не очікується значного позитивного зростання попиту через низьку покупноспроможність та невизначеність на ринку. Проте, після закінчення війни та встановлення економічної стабільності, поступово відбудеться відновлення покупельної активності. В цей час може спрацювати відкладений попит на ринку житла. Також, внутрішньо переміщені особи, які залишаться на своїх поточних місцях проживання, можуть розглянути можливість придбання власного житла. У довгостроковій перспективі відновлення попиту залежатиме від темпів макроекономічної стабілізації після війни.

Було виконано моделювання та прогнозування цін на будинки з використанням чотирьох методів машинного навчання: лінійна регресія, регуляризація, випадковий ліс та XGBoost. Після перевірки ефективності цих моделей на тестовому наборі даних та порівняння результатів між собою, було встановлено, що алгоритм XGBoost показує найкращі результати за метриками RMSE та MAPE, тоді як регуляризація показує найгірші результати. Виявлено, що багато вчених застосовують лінійну регресію для створення моделей та передбачення цін на нерухомість. З метою продемонструвати, що сучасні алгоритми краще впораються з цим завданням, було вирішено порівняти прогнози алгоритму лінійно регресії та XGBoost.

Результати виявили, що лінійна модель завищила ціну будинку на 23432,6 долара, що може призвести до того, що будинок не буде купуватись протягом тривалого періоду часу. Продавець, крім того, не тільки не заробить грошей, але й витратить їх на підтримку стану будинку, його рекламу, або буде змушений встановити ціну нижче ринкової і втратить ще більше коштів. У випадку з XGBoost моделлю, ціна будинку виявилась на 2422,2 долара дешевшою, що означає, що продавець отримав би менше, ніж він міг би. Однак, така різниця між реальною і прогнозованою ціною не є критичною, і швидкий продаж забезпечує відсутність додаткових витрат. З результату видно, що модель XGBoost краще відображає дані і має вищу ймовірність прогнозувати правильну ціну.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. CFI Team. What is Real Estate?. Corporate Finance Institute (CFI). URL: https://corporatefinanceinstitute.com/resources/careers/jobs/real-estate/ (дата звернення: 26.04.2023)

2. Chen J. What Is Real Estate?. Investopedia. URL: https://www.investopedia.com/terms/r/realestate.asp#:~:text=There%20are%20five%20main%20categories,estate%20investment%20trust%20(REIT) (дата звернення: 26.04.2023)

3. Glickman E. A. Principles of Real Estate Finance. An Introduction to Real Estate Finance. URL: https://www.sciencedirect.com/science/article/pii/B9780123786265020015 (дата звернення: 26.04.2023)

4. What Does the Housing Market Mean? My Accounting Course. URL: https://www.myaccountingcourse.com/accounting-dictionary/housing-market (дата звернення: 26.04.2023)

5. Галаган Д. В. Ринок нерухомості як різновид інвестиційного рин-ку. Теорія інвестицій. 2010. УДК 311.3:336. С.19-21. (дата звернення: 27.04.2023)

6. Youds C. What is the housing market?. The Sloman Economics News Site. URL: https://pearsonblog.campaignserver.co.uk/supply-and-demand-the-housing-market/ (дата звернення: 27.04.2023)

7. Що є об’єктом житлової нерухомості? Uteka.ua. URL: https://uteka.ua/publication/news-14-ezhednevnyj-buxgalterskij-obzor-39-chto-yavlyaetsya-obektom-zhiloj-nedvizhimosti (дата звернення: 27.04.2023)

8. Cooper R., John A.. Macroeconomics: Theory through Applications. 2011. 524p. (дата звернення: 28.04.2023)

9. Pham Duc Trung, Nguyen Gia Trung Quan. Factors affecting the price of the real estate: A case og Ho Chi Minck city. British Journal of Marketing Studies (BJMS). 2019. Vol. 7, Issue 6, P.35-45. (дата звернення: 28.04.2023)

10. Chandler T. Effect on the property market. Tracey Chandler - Buyers Agent. URL: https://buyersagent-sydney.com.au/12-factors-that-affect-property-prices/ (дата звернення: 30.04.2023)

11. Whitten R. 11 factors that affect property value. Finder. URL: https://www.finder.com.au/what-influences-a-propertys-value (дата звернення: 30.04.2023)

12. What factors affect property value?. Loans.com.au URL: https://www.loans.com.au/blog/10-factors-that-affect-property-value (дата звернення: 30.04.2023)

13. Gomez J. The most important factors that influence. Opendoor. URL: https://www.opendoor.com/w/blog/factors-that-influence-home-value (дата звернення: 30.04.2023)

14. Оцінка та управління нерухомістю: навчальний посібник / [В. Р. Кучеренко, М. А. Заєць, О. В. Захарченко, Н. В. Сментина, В. О. Улибіна]. – Одеса: Видавництво ТОВ «Лерадрук», 2013. – 272 с. (дата звернення: 30.04.2023)

15. Factors that affect property value. upside.com.au. URL: https://upside.com.au/articles/selling-your-property/selling-guide/9-surprising-factors-affect-home-value (дата звернення: 30.04.2023)

16. Bartsch C. 10 From the Global Stage to Your Backyard. HomeLight Blog. URL: https://www.homelight.com/blog/real-estate-property-value/ (дата звернення: 30.04.2023)

17. Nguyen J. The main factors that affect the real estate market. Investopedia. URL: https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp (дата звернення: 30.04.2023)

18. Що впливає на вартість нерухомого майна в Україні. Всі Новобудови. URL: https://vn.com.ua/ua/news/kakie-faktory-vlijajut-na-formirovanie-tseny-za-1-m2-v-zhk-i-kg (дата звернення: 30.04.2023)

19. Кузьміч О.Й., Іванова В.О. Аналіз деяких факторів, які впливають на вартість нерухомості. Містобудування та територіальне планування. УДК 528,48. С.251-257. (дата звернення: 30.04.2023)

20. Дослідження ринку нерухомості під час війни, Forbes. URL: https://forbes.ua/money/budivnitstva-ne-zapuskayutsya-popitu-nemae-ale-tsini-ne-padayut-doslidzhennya-rinku-nerukhomosti-pid-chas-viyni-vid-ernst-amp-young-13032023-12337 (дата звернення: 30.04.2023)

21. Тарасовський Ю. CEO забудовника «Ковальська» розповів, як обвалився їхній бізнес. Forbes. URL: https://forbes.ua/news/z-pochatku-viyni-prodali-tri-desyatki-kvartir-seo-zabudovnika-kovalska-rozpoviv-yak-obvalivsya-ikh-biznes-29112022-10106 (дата звернення: 30.04.2023)

22. Заржевська С. Забудовники повернули у продаж десятки новобудов, але чи є на них попит? URL: https://forbes.ua/money/yaki-tsini-na-nerukhomist-naspravdi-zabudovniki-povernuli-u-prodazh-desyatki-novobudov-ale-chi-e-na-nikh-popit-20092022-8448 (дата звернення: 30.04.2023)

23. Соломаха О. Як ціни на оренду житла стрибали у воєнний рік та що з ними зараз. Дослідження. Forbes. URL: https://forbes.ua/money/8000-grn-za-odnokimnatu-v-kievi-ta-16-000-u-lvovi-yak-tsini-na-orendu-zhitla-stribali-u-voenniy-rik-ta-shcho-z-nimi-zaraz-doslidzhennya-08032023-12232 (дата звернення: 30.04.2023)

24. Дослідження ринку нерухомості під час війни від EY. EY Ukraine. URL: https://www.ey.com/uk_ua/news/ey-ukraine-in-media/doslidzhennya-rynku-nerukhomosti-pid-chas-viyny-vid-ey (дата звернення: 30.04.2023)

25. Stobierski T. What is Statistical Modeling For Data Analy-sis?. Northeastern University Graduate Programs. URL:

https://www.northeastern.edu/graduate/blog/statistical-modeling-for-data-analysis/ (дата звернення: 30.04.2023)

26. What is Statistical Modeling? Definition and FAQs | OmniS-ci. Accelerated Analytics Platform | OmniSci. URL: https://www.omnisci.com/technical-glossary/statistical-modeling (дата звернення: 02.05.2023)

27. Yan, Xin. Linear regression analysis : theory and computing / by Xin Yan & Xiaogang Su. World Scientific Publishing Co. Pte. Ltd., 2009. 349 p. (дата звернення: 02.05.2023)

28. 5 Types of Regression Analysis And When To Use Them ｜Appier. Appier. URL: https://www.appier.com/blog/5-types-of-regression-analysis-and-when-to-use-them/ (дата звернення: 02.05.2023)

29. Helmuth Spath. Mathematical Algorithm for Linear Regression. Academic Press, Inc., 1992. 335 p. (дата звернення: 03.05.2023)

30. Vadapalli P. 6 Types of Regression Models in Machine Learning You Should Know About | upGrad blog. upGrad blog. URL: https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/ (дата звернення: 03.05.2023)

31. Brownlee J. Linear Regression for Machine Learning. Machine Learning Mastery. URL: https://machinelearningmastery.com/linear-regression-for-machine-learning/ (дата звернення: 03.05.2023)

32. Kenton W. How the Least Squares Method Works. Investopedia. URL: https://www.investopedia.com/terms/l/least-squares-method.asp#:~:text=The%20least%20squares%20method%20is%20a%20statistical%20procedure%20to%20find,the%20behavior%20of%20dependent%20variables (дата звернення: 03.05.2023)

33. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого – М.Ж ДМК Пресс, 2017. – 456 с. (дата звернення: 03.05.2023)

34. Простая линейная регрессия в EXCEL. Примеры и описание. Microsoft Excel в примерах и задачах. URL: https://excel2.ru/articles/prostaya-lineynaya-regressiya-v-ms-excel#standart-error (дата звернення: 03.05.2023)

35. Hayes A. Null Hypothesis Definition. Investopedia. URL: https://www.investopedia.com/terms/n/null_hypothesis.asp (дата звернення: 03.05.2023)

36. McLeod D. S. P-Value and Statistical Significance | Simply Psychol-ogy. Study Guides for Psychology Students - Simply Psychology. URL: https://www.simplypsychology.org/p-value.html (дата звернення: 03.05.2023)

37. Contributors to Wikimedia projects. Mean squared error - Wikipe-dia. Wikipedia, the free encyclopedia. URL: https://en.wikipedia.org/wiki/Mean_squared_error (дата звернення: 03.05.2023)

38. Использование критерия Фишера для проверки значимости ре-грессионной модели. Chem-astu. URL: https://www.chem-astu.ru/science/reference/fischer.html (дата звернення: 05.05.2023)

39. Tripathi M. Underfitting and Overfitting in Machine Learning. Data Science Articles and Whitepapers | Data Science Awards | Data Science Consul-tancy. URL: https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning (дата звернення: 05.05.2023)

40. Jerome Friedman, Trevor Hastie, Robert Tibshirani. The Elements of Statistical Learning Data Mining, Inference and Prediction. 2008. 809 p. (дата звернення: 05.05.2023)

41. Mishra M. REGULARIZATION: An important concept in Machine Learning. Medium. URL: https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea (дата звернення: 05.05.2023)

42. Gupta P. Regularization in Machine Learning. Medium. URL: https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a (дата звернення: 05.05.2023)

43. Jain S. Regularization Techniques | Regularization In Deep Learn-ing. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/ (дата звернення: 05.05.2023)

44. Cross-validation: evaluating estimator performance – scikit-learn 0.24.2 documentation. scikit-learn: machine learning in Python – scikit-learn 0.16.1 documentation. URL: https://scikit-learn.org/stable/modules/cross_validation.html (дата звернення: 05.05.2023)

45. Deol G. An Introduction to Ridge, Lasso, and Elastic Net Regres-sion. Hacker Noon. URL: https://hackernoon.com/an-introduction-to-ridge-lasso-and-elastic-net-regression-cca60b4b934f (дата звернення: 07.05.2023)

46. Decision Tree Tutorials & Notes | Machine Learning | Hack-erEarth. HackerEarth. URL: https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/ (дата звернення: 07.05.2023)

47. Decision Tree Regression. Data Mining Map. URL: https://www.saedsayad.com/decision_tree_reg.htm (дата звернення: 07.05.2023)

48. Cravit R. What is a Decision Tree and How to Make One [Templates + Examples] - Venngage. Venngage. URL: https://venngage.com/blog/what-is-a-decision-tree/ (дата звернення: 07.05.2023)

49. Machine Learning Decision Tree Classification Algorithm - Ja-vatpoint. www.javatpoint.com. URL: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (дата звернення: 07.05.2023)

50. R A. Regrssion in Decision Tree – A Step by Step CART (Classifica-tion And Regression Tree). Medium. URL: https://arifromadhan19.medium.com/regrssion-in-decision-tree-a-step-by-step-cart-classification-and-regression-tree-196c6ac9711e (дата звернення: 07.05.2023)

51. Machine Learning Random Forest Algorithm - Ja-vatpoint. www.javatpoint.com. URL: https://www.javatpoint.com/machine-learning-random-forest-algorithm (дата звернення: 07.05.2023)

52. Bagging (Bootstrap Aggregation) - Overview, How It Works, Ad-vantages. Corporate Finance Institute (CFI). URL: https://corporatefinanceinstitute.com/resources/knowledge/other/bagging-bootstrap-aggregation/ (дата звернення: 07.05.2023)

53. Mwiti D. Random Forest Regression: When Does It Fail and Why? - neptune.ai. neptune.ai. URL: https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why (дата звернення: 07.05.2023)

54. Cirillo A. R Data Mining. O'Reilly Online Learning. URL: https://www.oreilly.com/library/view/r-data-mining/9781787124462/85d1c737-e76f-4c79-80f9-b5eddff8444a.xhtml (дата звернення: 07.05.2023)

55. Chakure A. Random Forest and Its Implementation. Medium. URL: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f (дата звернення: 07.05.2023)

56. Cheng L. Basic Ensemble Learning (Random Forest, AdaBoost, Gra-dient Boosting)- Step by Step Explained. Medium. URL: https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725 (дата звернення: 07.05.2023)

57. XGBoost Algorithm | XGBoost In Machine Learning. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ (дата звернення: 08.05.2023)

58. Harode R. XGBoost: A Deep Dive into Boosting. Medium. URL: https://medium.com/sfu-cspmp/xgboost-a-deep-dive-into-boosting-f06c9c41349 (дата звернення: 08.05.2023)

59. Samudrala A. Unveiling Mathematics Behind XGBoost - KDnug-gets. KDnuggets. URL: https://www.kdnuggets.com/2018/08/unveiling-mathematics-behind-xgboost.html (дата звернення: 08.05.2023)

60. Fabien M. Gradient Boosting Regression. Welcome -. URL: https://maelfabien.github.io/machinelearning/GradientBoost/ (дата звернення: 08.05.2023)

61. Understanding XGBoost Algorithm In Detail. Analytics India Maga-zine. URL: https://analyticsindiamag.com/xgboost-internal-working-to-make-decision-trees-and-deduce-predictions/ (дата звернення: 08.05.2023)

62. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. URL: https://arxiv.org/pdf/1603.02754.pdf (дата звернення: 10.05.2023)

63. House Sales in King County, USA. Kaggle. URL: https://www.kaggle.com/harlfoxem/housesalesprediction (дата звернення: 10.05.2023)

64. Shivhare P. Predicting King County House Prices. SlideShare – a Scribd company. URL: https://www.slideshare.net/PawanShivhare1/predicting-king-county-house-prices (дата звернення: 10.05.2023)

65. Parcel Record Assessor extract table. URL: https://www5.kingcounty.gov/sdc/FGDCDocs/PARCEL_EXTR_faq.htm (дата звернення: 10.05.2023)

66. Glossary for Improved Sales. King County, Washington - King County. URL: https://kingcounty.gov/Assessor/Reports/ArchivedAreaReports/~/media/Assessor/AreaReports/AreaReportGlossary.ashx (дата звернення: 10.05.2023)

67. Bath-room. Wikidwelling. URL: https://wikidwelling.fandom.com/wiki/Bathroom (дата звернення: 10.05.2023)

68. Best Time of the Year to Buy a House. Better Mortgage Resources. URL: https://better.com/content/best-time-of-the-year-to-buy-a-house/ (дата звернення: 10.05.2023)

69. Department of Assessments. Medina/ Clyde Hill/ Hunts Point. Residential Revalue for 2019 Assessment Roll. 2019. 1-36 pp. (дата звернення: 10.05.2023)

70. Department of Assessments. Mercer Island. Residential Revalue for 2020 Assessment Roll. 2020. 1-61 pp. (дата звернення: 10.05.2023)

71. Taylor C. Detect the Presence of Outliers with the Interquartile Range Rule. ThoughtCo. URL: https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244 (дата звернення: 10.05.2023)

72. Bevans R. Linear Regression in R | An Easy Step-by-Step Guide. Scribbr. URL: https://www.scribbr.com/statistics/linear-regression-in-r/ (дата звернення: 10.05.2023)

73. The detrimental consequences of setting an excessively high price for your home. Duet Property. URL: https://www.duetproperty.com.au/2020/02/why-overpricing-your-home-is-a-bad-idea/#:~:text=The%20first%20few%20days%20after,buyers,%20significantly%20decreasing%20buyer%20demand (дата звернення: 10.05.2023)

74. Zillow: Real Estate, Apartments, Mortgages & Home Values. URL: https://www.zillow.com/ (дата звернення: 12.05.2023)

75. Redfin: Real Estate, Homes for Sale, MLS Listings, Agents. URL: https://www.redfin.com (дата звернення: 12.05.2023)

76. Trulia: Real Estate Listings, Homes For Sale, Housing Data. URL: https://www.trulia.com (дата звернення: 12.05.2023)

77. Compass Real Estate. URL: https://www.compass.com/ (дата звернення: 12.05.2023)

78. Homes for Sale, Apartments & Houses for Rent. URL: https://www.realtor.com (дата звернення: 12.05.2023)

ДОДАТКИ

Додаток А. Значення індексів Grade (показник рівня відповідності будівництва та дизайну)

Показник Grade	Опис
1-3	Будинок не відповідає мінімальним будівельним стандартам; зазвичай має примітивну будівлю без зручностей.
4	Головним чином старе будівництво низької якості. Будинок не відповідає нормам.
5	Низькі витрати на будівництво і обробку з обмеженою якістю. Будинок невеликий і має простий дизайн.
6	Найнижчий рівень відповідно до сучасних будівельних норм. Використання низькоякісних матеріалів та простих конструкцій.
7	Середня оцінка для конструкції та дизайну. Зазвичай це характерно для досить старих будинків.
8	Цей рівень трохи перевищує середній стандарт щодо конструкції та дизайну. У таких будинках, зазвичай, використовуються вищі якісні матеріали для зовнішньої та внутрішньої обробки.
9	Цей рівень пропонує кращий архітектурний дизайн з розширеним виконанням екстер'єру та інтер'єру, відзначається високою якістю.
10	Будинки цього рівня зазвичай мають високу якість. Оздоблення стало ще кращим, більше уваги приділяється якісному дизайну планування поверхів та збільшенню площі.
11	Цей рівень пропонує індивідуальний дизайн та використання вищої якості обробки, а також додаткові зручності, відмінну сантехніку та інші розкішні опції.
12	Тут ви знайдете індивідуальний дизайн та висококваліфікованих будівельників. Усі використані матеріали є найвищої якості, а будинки обладнані всіма зручностями.
13	Ці будинки зазвичай спроектовані та побудовані на замовлення, що наближає їх до рівня особняків. Вони відрізняються великою кількістю вбудованих шаф, використанням дерев'яної обробки та мармуру, а також мають великі двері.

```
Call:
lm(formula = price ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-660734  -49264   -2606   42898  568210

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     301109.7    36708.5   8.203 2.55e-16 ***
date              9723.5      721.3  13.480  < 2e-16 ***
bedrooms1       -51037.2    30133.7  -1.694 0.090346 .
bedrooms2       -58320.0    29286.1  -1.991 0.046457 *
bedrooms3       -61337.1    29257.3  -2.096 0.036058 *
bedrooms4       -61168.8    29301.7  -2.088 0.036856 *
bedrooms5       -76415.1    29443.0  -2.595 0.009459 **
bedrooms6       -94205.0    30204.4  -3.119 0.001819 **
bedrooms7      -163566.8    34705.1  -4.713 2.46e-06 ***
bedrooms8      -127903.9    42341.4  -3.021 0.002526 **
bedrooms9       -67938.2    67575.6  -1.005 0.314738
bedrooms10     -174140.1    67416.8  -2.583 0.009803 **
bathrooms         8226.1     1269.4   6.481 9.44e-11 ***
sqft_living      55052.4     2794.4  19.701  < 2e-16 ***
sqft_lot         11632.9      756.9  15.370  < 2e-16 ***
floors1.5         2944.2     2881.1   1.022 0.306838
floors2         -16220.0     2490.4  -6.513 7.61e-11 ***
floors2.5       -15675.5    10053.7  -1.559 0.118977
floors3         -57422.3     5408.0 -10.618  < 2e-16 ***
floors3.5       -12639.5    43676.2  -0.289 0.772287
waterfront1     149340.3    15494.9   9.638  < 2e-16 ***
view1            63978.8     6343.1  10.086  < 2e-16 ***
view2            66070.4     3812.5  17.330  < 2e-16 ***
view3            93559.6     5729.4  16.330  < 2e-16 ***
view4           166212.7     9653.3  17.218  < 2e-16 ***
condition2       81815.1    22253.0   3.677 0.000237 ***
condition3       99291.4    20990.9   4.730 2.26e-06 ***
condition4      118395.7    20991.6   5.640 1.73e-08 ***
condition5      151883.0    21101.3   7.198 6.43e-13 ***
grade            49811.9     1253.2  39.748  < 2e-16 ***
sqft_above       41422.0     2925.8  14.158  < 2e-16 ***
sqft_basement1    6990.0     2932.4   2.384 0.017150 *
yr_renovated1    35618.0     3999.5   8.906  < 2e-16 ***
age              11970.7     1357.2   8.820  < 2e-16 ***
zipcode98002      1267.1     9461.1   0.134 0.893458
zipcode98003     -9910.1     8447.2  -1.173 0.240740
zipcode98004    466799.8    16101.6  28.991  < 2e-16 ***
zipcode98005    289926.5    16442.8  17.632  < 2e-16 ***
zipcode98006    244172.9    13651.1  17.887  < 2e-16 ***
zipcode98007    210697.5    17047.7  12.359  < 2e-16 ***
zipcode98008    191527.7    16220.4  11.808  < 2e-16 ***
zipcode98010     92210.6    13946.1   6.612 3.93e-11 ***
zipcode98011     77572.7    20865.1   3.718 0.000202 ***
zipcode98014     62562.0    23410.3   2.672 0.007539 **
zipcode98019     37915.6    22732.6   1.668 0.095358 .
zipcode98022     20213.5    12733.5   1.587 0.112438
zipcode98023    -26477.8     7658.2  -3.457 0.000547 ***
zipcode98024    121859.8    20574.2   5.923 3.24e-09 ***
zipcode98027    172418.1    14032.3  12.287  < 2e-16 ***
zipcode98028     56545.2    20143.4   2.807 0.005005 **
zipcode98029    198625.2    15906.6  12.487  < 2e-16 ***
zipcode98030     -2341.2     9191.7  -0.255 0.798948
zipcode98031      1328.3     9600.0   0.138 0.889957
zipcode98032    -16674.1    11185.6  -1.491 0.136068
zipcode98033    262988.3    17579.7  14.960  < 2e-16 ***
zipcode98034    115342.9    18629.1   6.192 6.12e-10 ***
zipcode98038     44064.0    10636.3   4.143 3.45e-05 ***
zipcode98039    573387.4    45145.8  12.701  < 2e-16 ***
zipcode98040    371421.2    14475.4  25.659  < 2e-16 ***
zipcode98042      8372.4     8894.8   0.941 0.346583
zipcode98045    102870.0    19944.4   5.158 2.53e-07 ***
zipcode98052    201596.6    17797.0  11.328  < 2e-16 ***
zipcode98053    187672.0    19281.8   9.733  < 2e-16 ***
zipcode98055     16374.1    10543.2   1.553 0.120433
zipcode98056     69118.3    11634.1   5.941 2.90e-09 ***
zipcode98058     18547.2    10171.8   1.823 0.068266 .
zipcode98059     78959.2    11426.6   6.910 5.05e-12 ***
zipcode98065    114804.2    18276.7   6.281 3.45e-10 ***
zipcode98070     48640.7    13694.8   3.552 0.000384 ***
zipcode98072    103647.8    20816.9   4.979 6.47e-07 ***
zipcode98074    179675.4    16996.8  10.571  < 2e-16 ***
zipcode98075    207007.1    16457.1  12.579  < 2e-16 ***
zipcode98077    116900.1    21778.2   5.368 8.10e-08 ***
zipcode98092     -8143.6     8315.5  -0.979 0.327437
zipcode98102    345649.6    18387.8  18.798  < 2e-16 ***
zipcode98103    247431.1    16731.4  14.788  < 2e-16 ***
zipcode98105    312715.0    17559.4  17.809  < 2e-16 ***
zipcode98106     65117.0    12397.6   5.252 1.52e-07 ***
zipcode98107    244179.5    17232.2  14.170  < 2e-16 ***
zipcode98108     66666.5    13642.3   4.887 1.04e-06 ***
zipcode98109    344057.6    18065.4  19.045  < 2e-16 ***
zipcode98112    366535.4    16256.0  22.548  < 2e-16 ***
zipcode98115    240467.8    17007.6  14.139  < 2e-16 ***
zipcode98116    231279.9    13894.9  16.645  < 2e-16 ***
zipcode98117    231593.7    17167.7  13.490  < 2e-16 ***
zipcode98118    117071.1    12052.2   9.714  < 2e-16 ***
zipcode98119    331540.8    17140.3  19.343  < 2e-16 ***
zipcode98122    246547.1    14999.5  16.437  < 2e-16 ***
zipcode98125    119165.0    18359.2   6.491 8.82e-11 ***
zipcode98126    133973.9    12634.7  10.604  < 2e-16 ***
zipcode98133     66959.3    18891.6   3.544 0.000395 ***
zipcode98136    203410.5    12938.5  15.721  < 2e-16 ***
zipcode98144    186513.6    13899.2  13.419  < 2e-16 ***
zipcode98146     58420.0    11573.2   5.048 4.52e-07 ***
zipcode98148     24343.0    16190.3   1.504 0.132719
zipcode98155     53567.2    19738.8   2.714 0.006660 **
zipcode98166     60419.5    10677.8   5.658 1.56e-08 ***
zipcode98168     12807.9    11175.8   1.146 0.251798
zipcode98177    132414.7    19759.0   6.701 2.14e-11 ***
zipcode98178     16568.5    11542.2   1.435 0.151176
zipcode98188      5447.1    11484.7   0.474 0.635301
zipcode98198     -2425.0     8974.0  -0.270 0.786993
zipcode98199    287579.7    16523.5  17.404  < 2e-16 ***
lat              25402.6     5823.2   4.362 1.30e-05 ***
long             -7232.9     4456.7  -1.623 0.104631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85490 on 14221 degrees of freedom
Multiple R-squared:  0.8328,    Adjusted R-squared:  0.8316
F-statistic: 681.3 on 104 and 14221 DF,  p-value: < 2.2e-16
```

Рисунок Б.1 – Результат лінійної регресійної моделі

```
Call:
lm(formula = price ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-651249  -50157   -2833   43040  581603

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     304890.83   36875.33   8.268  < 2e-16 ***
bedrooms1       -49972.11   30325.51  -1.648 0.099404 .
bedrooms2       -56820.34   29471.17  -1.928 0.053875 .
bedrooms3       -59939.30   29442.09  -2.036 0.041785 *
bedrooms4       -59499.60   29486.60  -2.018 0.043625 *
bedrooms5       -74389.87   29628.32  -2.511 0.012058 *
bedrooms6       -92576.31   30394.96  -3.046 0.002325 **
bedrooms7      -162540.50   34923.67  -4.654 3.28e-06 ***
bedrooms8      -124972.86   42609.76  -2.933 0.003363 **
bedrooms9       -58006.52   68001.79  -0.853 0.393666
bedrooms10     -169465.19   67838.31  -2.498 0.012498 *
bathrooms         8060.94    1277.37   6.311 2.86e-10 ***
sqft_living      54664.37    2812.09  19.439  < 2e-16 ***
sqft_lot         11630.92     759.30  15.318  < 2e-16 ***
floors1.5         2771.04    2896.56   0.957 0.338753
floors2         -16000.58    2505.64  -6.386 1.76e-10 ***
floors2.5       -14970.52   10116.95  -1.480 0.138963
floors3         -56379.09    5439.27 -10.365  < 2e-16 ***
floors3.5        -6396.32   43951.96  -0.146 0.884295
waterfront1     149176.05   15591.59   9.568  < 2e-16 ***
view1            64906.01    6383.10  10.168  < 2e-16 ***
view2            66589.41    3836.65  17.356  < 2e-16 ***
view3            94047.40    5765.52  16.312  < 2e-16 ***
view4           166864.30    9714.60  17.177  < 2e-16 ***
condition2       82418.36   22394.56   3.680 0.000234 ***
condition3       99425.42   21124.28   4.707 2.54e-06 ***
condition4      117686.24   21124.90   5.571 2.58e-08 ***
condition5      149662.35   21234.62   7.048 1.90e-12 ***
grade            49505.20    1260.20  39.284  < 2e-16 ***
sqft_above       41727.40    2944.06  14.173  < 2e-16 ***
sqft_basement1    7498.84    2950.81   2.541 0.011055 *
yr_renovated1    33647.51    4022.10   8.366  < 2e-16 ***
age              12255.67    1364.49   8.982  < 2e-16 ***
zipcode98002        84.19    9343.91   0.009 0.992811
zipcode98003     -6081.58    8410.91  -0.723 0.469656
zipcode98004    458794.05   15843.63  28.958  < 2e-16 ***
zipcode98005    282536.16   15961.04  17.702  < 2e-16 ***
zipcode98006    235757.42   12875.88  18.310  < 2e-16 ***
zipcode98007    200641.32   16352.67  12.270  < 2e-16 ***
zipcode98008    180140.68   15181.14  11.866  < 2e-16 ***
zipcode98010     80398.53   11259.00   7.141 9.73e-13 ***
zipcode98011     69378.82   20645.68   3.360 0.000780 ***
zipcode98014     38134.42   18894.16   2.018 0.043577 *
zipcode98019     17717.45   19870.80   0.892 0.372605
zipcode98022      9054.17    9745.74   0.929 0.352885
zipcode98023    -20676.13    7138.07  -2.897 0.003778 **
zipcode98024    102609.64   16620.91   6.174 6.86e-10 ***
zipcode98027    157088.33   11663.89  13.468  < 2e-16 ***
zipcode98028     50366.19   20128.21   2.502 0.012351 *
zipcode98029    182002.40   13029.04  13.969  < 2e-16 ***
zipcode98030     -4988.51    8830.04  -0.565 0.572118
zipcode98031     -1947.24    9211.40  -0.211 0.832582
zipcode98032    -16071.64   11256.72  -1.428 0.153389
zipcode98033    254970.43   17218.22  14.808  < 2e-16 ***
zipcode98034    107603.17   18414.27   5.843 5.22e-09 ***
zipcode98038     32900.65    7420.20   4.434 9.32e-06 ***
zipcode98039    569481.83   45366.38  12.553  < 2e-16 ***
zipcode98040    366008.19   14371.11  25.468  < 2e-16 ***
zipcode98042      1549.46    7369.48   0.210 0.833473
zipcode98045     74283.17   11197.71   6.634 3.39e-11 ***
zipcode98052    188943.61   16834.00  11.224  < 2e-16 ***
zipcode98053    170756.66   17150.46   9.956  < 2e-16 ***
zipcode98055     13144.79   10258.37   1.281 0.200084
zipcode98056     63925.80   11167.84   5.724 1.06e-08 ***
zipcode98058     11882.77    9224.31   1.288 0.197697
zipcode98059     70324.47   10458.98   6.724 1.84e-11 ***
zipcode98065     91879.76   12363.28   7.432 1.13e-13 ***
zipcode98070     56826.78   12503.83   4.545 5.55e-06 ***
zipcode98072     93217.59   20093.10   4.639 3.53e-06 ***
zipcode98074    164015.28   14968.89  10.957  < 2e-16 ***
zipcode98075    192008.74   13991.32  13.723  < 2e-16 ***
zipcode98077    101140.17   20374.76   4.964 6.99e-07 ***
zipcode98092    -12677.49    7759.46  -1.634 0.102321
zipcode98102    343290.03   18500.24  18.556  < 2e-16 ***
zipcode98103    246857.39   16813.76  14.682  < 2e-16 ***
zipcode98105    311344.35   17664.26  17.626  < 2e-16 ***
zipcode98106     68176.23   12342.50   5.524 3.38e-08 ***
zipcode98107    246146.36   17237.51  14.280  < 2e-16 ***
zipcode98108     64619.08   13726.89   4.707 2.53e-06 ***
zipcode98109    344857.37   18139.88  19.011  < 2e-16 ***
zipcode98112    362819.86   16354.62  22.185  < 2e-16 ***
zipcode98115    237982.45   17110.29  13.909  < 2e-16 ***
zipcode98116    234984.04   13693.42  17.160  < 2e-16 ***
zipcode98117    232388.22   17165.59  13.538  < 2e-16 ***
zipcode98118    116168.38   12105.78   9.596  < 2e-16 ***
zipcode98119    331836.00   17160.73  19.337  < 2e-16 ***
zipcode98122    244911.73   15092.90  16.227  < 2e-16 ***
zipcode98125    115672.00   18467.14   6.264 3.87e-10 ***
zipcode98126    136445.02   12490.54  10.924  < 2e-16 ***
zipcode98133     66419.93   18997.85   3.496 0.000473 ***
zipcode98136    207128.02   12722.09  16.281  < 2e-16 ***
zipcode98144    185230.50   13986.63  13.243  < 2e-16 ***
zipcode98146     61823.02   11470.51   5.390 7.17e-08 ***
zipcode98148     25504.81   16237.27   1.571 0.116261
zipcode98155     50487.31   19855.22   2.543 0.011008 *
zipcode98166     63764.37   10576.55   6.029 1.69e-09 ***
zipcode98168     11252.00   11233.40   1.002 0.316527
zipcode98177    131424.33   19829.16   6.628 3.53e-11 ***
zipcode98178     13372.53   11518.75   1.161 0.245687
zipcode98188      5944.20   11556.07   0.514 0.606995
zipcode98198      -148.14    8965.26  -0.017 0.986817
zipcode98199    290245.80   16408.15  17.689  < 2e-16 ***
lat              27065.93    5838.48   4.636 3.59e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86030 on 14223 degrees of freedom
Multiple R-squared:  0.8307,    Adjusted R-squared:  0.8295
F-statistic: 684.1 on 102 and 14223 DF,  p-value: < 2.2e-16
```

Рисунок В.1 – Результат лінійної регресійної моделі без слабокорельованих показників

```
data <- kc_house_data
data$id = NULL
set.seed(123)
Q1 <- quantile(data$price, probs=.25)
Q3 <- quantile(data$price, probs=.75)
iqr = Q3-Q1
upper_limit = Q3 + (iqr*1.5)
lower_limit = Q1 - (iqr*1.5)
data$price[data$price > upper_limit]
data$price[data$price < lower_limit]
data <- subset(data, data$price > lower_limit & data$price < upper_limit)
data$yr_built = NULL
data$repair_age = NULL
boxplot(data$price ~ data$date, main = 'Price vs Date', col=c("blue","red"))
boxplot(data$price ~ data$bedrooms, main = 'Price vs Bedrooms', col=c("blue","red"))
print(subset(data, data$bedrooms > 10))
data <- data[data$bedrooms <= 10, ]
data$bedrooms = as.factor(data$bedrooms)
boxplot(data$price ~ data$bathrooms, main = 'Price vs Bathrooms', col=c("blue","red"))
boxplot(data$price ~ data$sqft_living, main = 'Price vs Sqft_living', col=c("blue","red"))
boxplot(data$price ~ data$sqft_lot, main = 'Price vs Sqft_lot', col=c("blue","red"))
boxplot(data$price ~ data$floors, main = 'Price vs floors', col=c("blue","red"))
data$floors = as.factor(data$floors)
```

```
boxplot(data$price ~ data$waterfront, main = 'Price vs waterfront', col=c("blue","red"))
data$waterfront = as.factor(data$waterfront)
boxplot(data$price ~ data$view, main = 'Price vs View', col=c("blue","red"))
data$view = as.factor(data$view)
boxplot(data$price ~ data$condition, main = 'Price vs Condition', col=c("blue","red"))
data$condition = as.factor(data$condition)
boxplot(data$price ~ data$grade, main = 'Price vs Grade', col=c("blue","red"))
boxplot(data$price ~ data$sqft_above, main = 'Price vs sqft_above', col=c("blue","red"))
boxplot(data$price ~ data$sqft_basement, main = 'Price vs Sqft_basement', col=c("blue","red"))
length(data$sqft_basement[data$sqft_basement == 0])
data$sqft_basement[data$sqft_basement != 0] = 1
data$sqft_basement = as.factor(data$sqft_basement)
boxplot(data$price ~ data$yr_renovated, main = 'Price vs yr_renovated', col=c("blue","red"))
length(data$yr_renovated[data$yr_renovated == 0])
data$yr_renovated[data$yr_renovated != 0] = 1
data$yr_renovated = as.factor(data$yr_renovated)
boxplot(data$price ~ data$age, main = 'Price vs age', col=c("blue","red"))
boxplot(data$price ~ data$zipcode, main = 'Price vs zipcode', col=c("blue","red"))
data$zipcode = as.factor(data$zipcode)
boxplot(data$price ~ data$lat, main = 'Price vs lat', col=c("blue","red"))
boxplot(data$price ~ data$long, main = 'Price vs long', col=c("blue","red"))
numeric_features = sapply(data[, -1], is.numeric)
numeric_features = c(FALSE, numeric_features) # Ціну не потрібно нормалізовувати
```

```
data[numeric_features] = sapply(data[numeric_features], scale)
set.seed(123)
n = nrow(data)
sample = sample(1:n, size = round(0.7*n), replace=FALSE)
train = data[sample ,]
test = data[-sample ,]
model_full = lm( price ~ .,data = train)
summary(model_full)
price_prediction_full <- predict(model_full, newdata = test)
res_full <- test$price-price_prediction_full
model_output_full <- cbind(test, price_prediction_full,res_full)
model_output_full <- cbind(test$price, price_prediction_full,res_full)
rsq_full <- cor(test$price, price_prediction_full)^2
rmse_full <- rmse(model_output_full$price,
model_output_full$price_prediction_full)
mae(model_output_full$price, model_output_full$price_prediction_full)
mape_full<-mape(model_output_full$price,
model_output_full$price_prediction_full)
smape(model_output_full$price, model_output_full$price_prediction_full)
mase(model_output_full$price, model_output_full$price_prediction_full)
step(lm(price ~.,data=train),direction="backward")
step(lm(price ~.,data=train),direction="forward")
step(lm(price ~.,data=train),direction="both")
data$date = NULL
data$long = NULL
model_nk = lm(formula = price ~ .,data = train)
summary(model_nk)
price_prediction <- predict(model_nk, newdata = test)
```

```
model_output <- cbind(test, price_prediction)
rsq_nk <- cor(test$price, price_prediction)^2 #0.831
rmse_nk <- rmse(model_output$price, model_output$price_prediction)
mae(model_output$price, model_output$price_prediction)
mape_nk <- mape(model_output$price,model_output$price_prediction)
smape(model_output$price, model_output$price_prediction)
mase(model_output$price, model_output$price_prediction)
rsq <- cbind("R-squared" = c(rsq_full, rsq_nk))
rownames(rsq) <- c("Full Regression","Regression NK" )
print(rsq)
rmse <- cbind("RMSE" = c(rmse_full, rmse_nk ))
rownames(rmse) <- c("Full Regression","Regression NK")
print(rmse)
mape <- cbind("MAPE" = c(mape_full, mape_nk ))
rownames(rmse) <- c("Full Regression","Regression NK")
print(mape)
result <- cbind(rsq,rmse,mape)
print(result)
```

```
library(tidyverse)
library(tidyr)
library(dplyr)
library(ggplot2)
library(car)
library(MASS)
library(caTools)
library(Metrics)
library(glmnet)
library(varImp)
data <- kc_house_data
data$id = NULL
set.seed(123)
Q1 <- quantile(data$price, probs=.25)
Q3 <- quantile(data$price, probs=.75)
iqr = Q3-Q1
upper_limit = Q3 + (iqr*1.5)
lower_limit = Q1 - (iqr*1.5)
data$price[data$price > upper_limit]
data$price[data$price < lower_limit]
data <- subset(data, data$price > lower_limit & data$price < upper_limit)
data$yr_built = NULL
data$repair_age = NULL
data <- data[data$bedrooms <= 10, ]
data$bedrooms = as.factor(data$bedrooms)
data$floors = as.factor(data$floors)
data$waterfront = as.factor(data$waterfront)
```

```
data$view = as.factor(data$view)
data$condition = as.factor(data$condition)
data$sqft_basement[data$sqft_basement != 0] = 1
data$sqft_basement = as.factor(data$sqft_basement)
data$yr_renovated[data$yr_renovated != 0] = 1
data$yr_renovated = as.factor(data$yr_renovated)
data$zipcode = as.factor(data$zipcode)
str(data)
numeric_features = sapply(data[, -1], is.numeric)
numeric_features = c(FALSE, numeric_features) # Ціну не потрібно нормалізовувати
data[numeric_features] = sapply(data[numeric_features], scale)
y <- data %>% dplyr::select(price) %>% data.matrix()
x <- data %>% dplyr::select(-price) %>% data.matrix()
n = nrow(data)
sample = sample(1:n, size = round(0.7*n), replace=FALSE)
x_train = x[sample ,]
x_test = x[-sample ,]
y_train = y[sample ,]
y_test = y[-sample ,]
data_ridge <- glmnet(
x = x_train,
y = y_train,
alpha = 0
)
plot(data_ridge, xvar = "lambda")
legend("bottomright", lwd = 2,col = 1:17, y.intersp = 0.3,legend = colnames(x_train),
cex = .9)
```

```
ridge <- cv.glmnet(x_train, y_train, type.measure="mse",
alpha=0, family="gaussian")
plot(ridge)
min(ridge$cvm)
ridge$lambda.min
ridge$cvm[ridge$lambda == ridge$lambda.1se]
ridge$lambda.1se
ridge_min <- glmnet(
x = x_train,
y = y_train,
alpha = 0
)
plot(ridge_min, xvar = "lambda")
legend("bottomright", lwd = 2,col = 1:17, y.intersp = 0.3,legend = colnames(x_train),
cex = .9)
abline(v = log(ridge$lambda.min), col = "red", lty = "dashed")
abline(v = log(ridge$lambda.1se), col = "red", lty = "dashed")
ridge_model <- glmnet(x_train, y_train, alpha = 0, lambda = ridge$lambda.1se,
standardize = TRUE)
ridge_predicted <- predict(ridge_model, s=ridge$lambda.1se, newx=x_test)
mse_ridge_cv <- mean((y_test - ridge_predicted)^2)
rsq_ridge_cv <- cor(y_test, ridge_predicted)^2 #0.691
data_lasso <- glmnet(
x = x_train,
y = y_train,
alpha = 1
)
plot(data_lasso, xvar = "lambda")
```

```
legend("bottomright", lwd = 2,col = 1:17, y.intersp = 0.3,legend = colnames(x_train), cex = .9)
lasso <- cv.glmnet(x_train, y_train, type.measure="mse",
alpha=1, family="gaussian")
plot(lasso)
min(lasso$cvm)
lasso$lambda.min
lasso$cvm[lasso$lambda == lasso$lambda.1se]
lasso$lambda.1se # лямбда для цього MSE 3482.313
lasso_min <- glmnet(
x = x_train,
y = y_train,
alpha = 1
)
plot(lasso_min, xvar = "lambda")
legend("bottomright", lwd = 2,col = 1:17, y.intersp = 0.3,legend = colnames(x_train), cex = .9)
abline(v = log(lasso$lambda.min), col = "red", lty = "dashed")
abline(v = log(lasso$lambda.1se), col = "red", lty = "dashed")
lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = lasso$lambda.1se, standardize = TRUE)
lasso_predicted <- predict(lasso_model, s=lasso$lambda.1se, newx=x_test)
mse_lasso_cv <- mean((y_test - lasso_predicted)^2) #13500403033
mse_lasso_cv
rsq_lasso_cv <- cor(y_test, lasso_predicted)^2 #0.69
rsq_lasso_cv
data_elastic_net <- glmnet(
x = x_train,
```

```
y = y_train,
alpha = 0.5
)
plot(data_elastic_net, xvar = "lambda")
legend("bottomright", lwd = 2,col = 1:17, y.intersp = 0.3,legend = colnames(x_train),
cex = .9)
elastic_net <- cv.glmnet(x_train, y_train, type.measure="mse",
alpha=0.5, family="gaussian")
plot(elastic_net)
min(elastic_net$cvm)
elastic_net$lambda.min
elastic_net$cvm[elastic_net$lambda == elastic_net$lambda.1se]
elastic_net$lambda.1se
elastic_net_min <- glmnet(
x = x_train,
y = y_train,
alpha = 0.5
)
plot(elastic_net_min, xvar = "lambda")
legend("bottomright", lwd = 2,col = 1:17, y.intersp = 0.3,legend = colnames(x_train),
cex = .9)
abline(v = log(elastic_net$lambda.min), col = "red", lty = "dashed")
abline(v = log(elastic_net$lambda.1se), col = "red", lty = "dashed")
en_model <- glmnet(x_train, y_train, alpha = 0.5, lambda = elastic_net$lambda.1se,
standardize = TRUE)
en_model
elastic_net_predicted <- predict(en_model, s=elastic_net$lambda.1se, newx=x_test)
mse_en_cv <- mean((y_test - elastic_net_predicted)^2) #13483769197
```

```
rsq_en_cv <- cor(y_test, elastic_net_predicted)^2 #0.691
rsq <- cbind("R-squared" = c(rsq_ridge_cv, rsq_lasso_cv, rsq_en_cv ))
rownames(rsq) <- c("ridge cross-validated", "lasso cross_validated", "elastic net cross_validated")
print(rsq)
mse <- cbind("Mse" = c(mse_ridge_cv, mse_lasso_cv, mse_en_cv ))
rownames(mse) <- c("ridge cross-validated", "lasso cross_validated", "elastic net cross_validated")
print(mse)
result <- cbind(rsq,mse)
print(result)
rmse(y_test,elastic_net_predicted)
mae(y_test,elastic_net_predicted)
mape(y_test,elastic_net_predicted)
smape(y_test,elastic_net_predicted)
mase(y_test,elastic_net_predicted)
```

```
library(tidyverse)
library(car)
library(MASS)
library(caTools)
library(Metrics)
library(cowplot)
library(randomForest)
library(rsample)
library(ranger)
library(caret)
library(h2o)
Sys.setenv(JAVA_HOME="D:/Java  13/zulu13.40.15-ca-jdk13.0.7-win_x64")  ##your own path of Java SE intalled
data <- kc_house_data
data$id = NULL
set.seed(123)
Q1 <- quantile(data$price, probs=.25)
Q3 <- quantile(data$price, probs=.75)
iqr = Q3-Q1
upper_limit = Q3 + (iqr*1.5)
lower_limit = Q1 - (iqr*1.5)
data$price[data$price > upper_limit]
data$price[data$price < lower_limit]
data <- subset(data, data$price > lower_limit & data$price < upper_limit)
data$yr_built = NULL
data$repair_age = NULL
data <- data[data$bedrooms <= 10, ]
```

```
data$bedrooms = as.factor(data$bedrooms)
data$floors = as.factor(data$floors)
data$waterfront = as.factor(data$waterfront)
data$view = as.factor(data$view)
data$condition = as.factor(data$condition)
data$sqft_basement[data$sqft_basement != 0] = 1
data$sqft_basement = as.factor(data$sqft_basement)
data$yr_renovated[data$yr_renovated != 0] = 1
data$yr_renovated = as.factor(data$yr_renovated)
str(data)
numeric_features = sapply(data[, -1], is.numeric)
numeric_features = c(FALSE, numeric_features) # Ціну не потрібно нормалізовувати
data[numeric_features] = sapply(data[numeric_features], scale)
n = nrow(data)
sample = sample(1:n, size = round(0.7*n), replace=FALSE)
train = data[sample ,]
test = data[-sample ,]
model_rf <- randomForest(price ~ ., data = train, proximity=TRUE )
plot(model_rf)
which.min (model_rf$mse)
sqrt(model_rf$mse [which.min (model_rf$ mse)])
randomForest_pred <- predict(model_rf, test)
rsq_rf <- cor(test$price, randomForest_pred)^2
rmse_rf <- rmse(test$price, randomForest_pred)
mae(test$price, randomForest_pred)
mape_rf <- mape(test$price, randomForest_pred)
smape(test$price, randomForest_pred)
```

```
mase(test$price, randomForest_pred)
set.seed(123)
h2o.no_progress()
h2o.init(max_mem_size = "5g")
y <- "price"
x <- setdiff(names(train), y)
train.h2o <- as.h2o(train)
test.h2o <- as.h2o(test)
hyper_grid.h2o <- list(
ntrees = seq(50, 500, by = 50),
mtries = seq(5, 35, by = 10),
max_depth = seq(20, 40, by = 5),
min_rows = seq(1, 5, by = 1),
nbins = seq(10, 30, by = 5),
sample_rate = c(.55, .632, .75)
)
search_criteria <- list(
strategy = "RandomDiscrete",
stopping_metric = "mse",
stopping_tolerance = 0.005,
stopping_rounds = 10,
max_runtime_secs = 30*60
)
random_grid <- h2o.grid(
algorithm = "randomForest",
grid_id = "rf_grid2",
x = x,
y = y,
```

```
training_frame = train.h2o,
hyper_params = hyper_grid.h2o,
search_criteria = search_criteria
)
grid_perf2 <- h2o.getGrid(
grid_id = "rf_grid2",
sort_by = "mse",
decreasing = FALSE
)
print(grid_perf2)
best_model_id <- grid_perf2@model_ids[[1]]
best_model <- h2o.getModel(best_model_id)
best_model <- h2o.randomForest(
x = x,
y = y,
training_frame = train.h2o,
ntrees = 450,
mtries = 15,
max_depth = 20,
min_rows = 1,
nbins = 30,
sample_rate = .55)
importance_matrix <- h2o.varimp(best_model)
h2o.varimp_plot(model = best_model , num_of_features = 17)
best_model_perf <- h2o.performance(model = best_model, newdata =test.h2o)
rsq_rf2 <- h2o.r2(best_model_perf)
rmse_rf2 <- h2o.mse(best_model_perf) %>% sqrt()
rsq <- cbind("R-squared" = c(rsq_rf, rsq_rf2))
```

```
rownames(rsq) <- c("Default RF","RF with Tuning HP" )
print(rsq)
rmse <- cbind("RMSE" = c(rmse_rf, rmse_rf2 ))
rownames(rmse) <- c("Default RF","RF with Tuning HP" )
print(rmse)
result <- cbind(rsq,rmse)
print(result)
```

```
library(tidyverse)
library(car)
library(MASS)
library(caTools)
library(Metrics)
library(rsample)
library(caret)
library(gbm)
library(xgboost)
data <- kc_house_data
data$id = NULL
set.seed(123)
Q1 <- quantile(data$price, probs=.25)
Q3 <- quantile(data$price, probs=.75)
iqr = Q3-Q1
upper_limit = Q3 + (iqr*1.5)
lower_limit = Q1 - (iqr*1.5)
data$price[data$price > upper_limit]
data$price[data$price < lower_limit]
data <- subset(data, data$price > lower_limit & data$price < upper_limit)
data$yr_built = NULL
data$repair_age = NULL
data <- data[data$bedrooms <= 10, ]
data$bedrooms = as.factor(data$bedrooms)
data$floors = as.factor(data$floors)
data$waterfront = as.factor(data$waterfront)
data$view = as.factor(data$view)
```

```
data$condition = as.factor(data$condition)
data$sqft_basement[data$sqft_basement != 0] = 1
data$sqft_basement = as.factor(data$sqft_basement)
data$yr_renovated[data$yr_renovated != 0] = 1
data$yr_renovated = as.factor(data$yr_renovated)
data$zipcode = as.factor(data$zipcode)
numeric_features = sapply(data[, -1], is.numeric)
numeric_features = c(FALSE, numeric_features) # Ціну не потрібно нормалізовувати
data[numeric_features] = sapply(data[numeric_features], scale)
y <- data %>% dplyr::select(price) %>% data.matrix()
x <- data %>% dplyr::select(-price) %>% data.matrix()
n = nrow(data)
sample = sample(1:n, size = round(0.7*n), replace=FALSE)
x_train = x[sample ,]
x_test = x[-sample ,]
y_train = y[sample ,]
y_test = y[-sample ,]
xgb_train = xgb.DMatrix(data = x_train, label = y_train)
xgb_test = xgb.DMatrix(data = x_test, label = y_test)
xgb.fit1 <- xgb.cv(
data = xgb_train,
label = xgb_test,
nrounds = 1000,
nfold = 10,
objective = "reg:linear",
verbose = 0
)
```

```
xgb.fit1$evaluation_log %>%
dplyr::summarise(
ntrees.train = which(train_rmse_mean == min(train_rmse_mean))[1],
rmse.train = min(train_rmse_mean),
ntrees.test = which(test_rmse_mean == min(test_rmse_mean))[1],
rmse.test = min(test_rmse_mean),
)
ggplot(xgb.fit1$evaluation_log) +
geom_line(aes(iter, train_rmse_mean), color = "red") +
geom_line(aes(iter, test_rmse_mean), color = "blue")
hyper_grid <- expand.grid(
eta = c(.01, .05, .1, .3),
gama = c(0, .25, 1),
lambda = c(0, .5, 5),
max_depth = c(1,3,5,7),
min_child_weight = c(1,3,5,7),
subsample = (.65),
colsample_bytree = c(.75, .9, 1),
optimal_trees = 0,
min_RMSE = 0
)
for(i in 1:nrow(hyper_grid)) {
params <- list(
eta = hyper_grid$eta[i],
max_depth = hyper_grid$max_depth[i],
min_child_weight = hyper_grid$min_child_weight[i],
subsample = hyper_grid$subsample[i],
colsample_bytree = hyper_grid$colsample_bytree[i]
```

```
)
xgb.tune <- xgb.cv(
params = params,
data = x_train,
label = y_train,
nrounds = 5000,
nfold = 10,
objective = "reg:linear",
verbose = 0,
early_stopping_rounds = 10
)
hyper_grid$optimal_trees[i] <- which.min(xgb.tune$evaluation_log$test_rmse_mean)
hyper_grid$min_RMSE[i] <- min(xgb.tune$evaluation_log$test_rmse_mean)
}
hyper_grid %>%
dplyr::arrange(min_RMSE) %>%
head(10)
params <- list(
eta = 0.01,
gama = 0.25,
lambda = 0.5,
max_depth = 7,
min_child_weight = 1,
subsample = 0.65,
colsample_bytree = 0.75
)
xgb.fit.final <- xgboost(
params = params,
```

```
data = x_train,
label = y_train,
nrounds = 2018,
objective = "reg:linear",
verbose = 0
)
importance_matrix <- xgb.importance(model = xgb.fit.final)
xgb.plot.importance(importance_matrix, top_n = 17, measure = "Gain")
pred <- predict(xgb.fit.final, x_test)
res_xgb <- y_test-pred
model_xgb_full <- cbind(y_test, pred,res_xgb)
model_xgb_full <- cbind(y_test, x_test, pred,res_xgb)
rsq_xgb <- cor(y_test, pred)^2
rmse(pred, y_test)
mae(pred, y_test)
mape(pred, y_test)
smape(pred, y_test)
mase(pred, y_test)
```