Issue 14, 2023

https://doi.org/10.23939/sisn2024.15.114

УДК 004.8

ПРОГНОЗУВАННЯ ВАРТОСТІ НЕРУХОМОСТІ З ВИКОРИСТАННЯМ ЗАСОБІВ МАШИННОГО НАВЧАННЯ

Верес Олег[1], Шимоняк Андрій[2]

Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж, Львів, Україна
[1] Oleh.M.Veres@lpnu.ua, ORCID 0000-0001-9149-4752
[2] andrii.shymoniak.msaad.2022@lpnu.ua, ORCID 0009-0003-7068-5818

Коректна оцінка вартості нерухомості відіграє вирішальну роль у купівлі та продажу. Вартість визначається різними факторами, такими як розташування, площа кухні та кімнат, стан, рік забудови, зручності, інфраструктура поблизу, тренди розвитку району, ринкові тенденції, та багатьма іншими. Розроблена модель допоможе продавцям отримати оцінку їхньої нерухомості за внесеними параметрами, що може слугувати відправною точкою для встановлення кінцевої вартості. Обчислення вартості нерухомості історично ґрунтувалось переважно на методі аналізу даних вручну та суб'єктивних оцінках, що часто призводило до помилок і затримок. Застосування алгоритмів машинного навчання для вирішення цієї проблеми виявилось ефективним, оскільки має низку переваг над методом оцінки вручну, а саме: високий рівень точності, запобігання суб'єктивності та упередженості в оцінках, ефективність у часі, зниження витрат, використання геопросторових даних і обґрунтування результатів. Процес створення моделі машинного навчання умовно розділено на чотири етапи, які містять збирання даних, їх фільтрування, оброблення, доповнення, розподіл на різні вибірки та тренування моделі на основі цих даних. Прийнято рішення використати одразу декілька алгоритмів регресії для побудови моделей машинного навчання, щоб порівняти результати та вибрати алгоритм, який найкраще підходить для вирішення поставленого завдання. Розглянуто найпопулярніші алгоритми регресії, коротко описано принцип їх роботи, а також метрики, за допомогою яких можна оцінити якість прогнозованих значень моделей. Для тестування алгоритмів лінійної регресії, дерева прийняття рішень, методу найближчого сусіда, методу опорних векторів і "випадкового лісу" застосовано стандартні параметри. Вибрано коефіцієнт детермінації R-квадрат як основну метрику. Порівняння коефіцієнтів детермінації отриманих результатів показало, що найкращий результат в алгоритму "випадковий ліс". Підібравши вручну гіперпараметри для цього алгоритму, досягнуто середнього значення абсолютної похибки прогнозованого значення, що становить 8,49 %, а медіани 1,9 %. Побудована модель відповідає встановленим вимогам якості та готова до реалізації у інформаційній системі прогнозування вартості нерухомості.

Ключові слова: аналіз даних; машинне навчання; регресія.

Вступ

Операції з нерухомістю передбачають залучення великих сум коштів, а отже, такі рішення потрібно приймати на основі релевантних даних. Коректна оцінка вартості нерухомості відіграє вирішальну роль у процесі купівлі та продажу. Вартість визначається різними факторами, такими як розташування, площа кухні та кімнат, стан, рік забудови, зручності, інфраструктура поблизу,

тренди розвитку району, ринкові тенденції, та багатьма іншими. Нерухомість зі завищеною вартістю може довго залишатись на ринку без уваги, тоді як продаж за заниженою вартістю спричиняє значні збитки для продавця. Важливо визначити точну оцінку, щоб приймати обґрунтовані рішення та уникнути помилок, що можуть призвести до фінансових втрат [1].

Проте вартість послуг фахівців нерідко висока. Крім того, у випадку покупки варто зважати на обмеження людських можливостей, особливо коли йдеться про великий обсяг даних. Треба врахувати, що пошук нерухомості може тривати довго, а іноді таке питання є терміновим. Отже, актуальним є завдання розроблення інформаційної системи, яка допоможе задовольнити потреби як продавців, так і покупців.

Реалізація цього проєкту дасть змогу спростити та пришвидшити пошук оптимальної нерухомості для покупки, а також процес її оцінки для подальшого продажу. Продавці нерухомості матимуть змогу швидко оцінити вартість їхнього майна залежно від параметрів, розташування та поточного стану ринку, а покупці – отримати список рекомендованої нерухомості, вартість якої встановлена відповідно до її реальних характеристик або є нижчою за ринкову. Проєкт спрямований на створення зручного та доступного інструменту, що надасть переваги для усіх зацікавлених сторін, враховуючи інвесторів, покупців, продавців та агентів нерухомості.

Постановка проблеми

У житті кожної людини настає момент, коли питання покупки чи продажу нерухомості набуває актуальності. Зазвичай така подія є надзвичайно важливою, оскільки передбачає залучення великих коштів та стається лише кілька разів у житті, тому потребує обдуманого та обґрунтованого підходу. Однак багато людей стикаються з обмеженими можливостями доступу до об’єктивної інформації щодо вартості нерухомості та тенденцій розвитку ринку. Тому наявність додатків, що допоможуть полегшити, пришвидшити, здешевити та раціоналізувати цей процес, завжди актуальна. З іншого боку, покупку нерухомості можна розглядати як різновид інвестицій, для яких важливі також вищезгадані швидкість, простота та раціональність.

Сьогодні наявні різноманітні вебплатформи та сервіси для пошуку нерухомості, але вони надають обмежену інформацію, оскільки не відображають її оціночну вартість, а також не завжди містять відповідність між вказаною продавцем ціною та реальними характеристиками нерухомості. Відсутні адекватні рекомендаційні системи, які враховували б та відображали б комплексну інформацію для покупців та продавців.

Визначення вартості нерухомості історично ґрунтувалось переважно на методі аналізу даних вручну та суб’єктивних оцінках, що часто призводило до помилок і затримок. Застосування алгоритмів машинного навчання для вирішення цієї проблеми виявилось ефективним, оскільки має низку переваг над методом оцінювання вручну, а саме: високий рівень точності, уникнення суб’єктивності та упередженості в оцінках, ефективність у часі, зниження витрат, використання геопросторових даних і обґрунтування результатів [2].

Отже, застосування методів машинного навчання для прогнозування вартості об’єктів нерухомості актуальне, оскільки забезпечить прозорість ринку нерухомості, дасть змогу зменшити витрати на послуги ріелторів та агентів з нерухомості, а користувачам – ефективніше приймати рішення стосовно покупки чи продажу.

Аналіз останніх досліджень та публікацій

Використання машинного навчання для прогнозування вартості нерухомості уможливило досягнення нечуваних раніше показників точності, ефективності та прозорості. Більше не треба повністю покладатися на судження експертів у цій галузі, яким часто доводилося працювати з величезними обсягами даних і встановлювати складні критерії для визначення цін на нерухомість.

Алгоритми штучного інтелекту працюють з величезною кількістю даних, аналізуючи їх значно швидше, ніж це фізично може зробити будь-яка людина, та перетворюючи їх на корисну інфор-

мацію, яка надалі стане основою для оцінок. Це надає змогу забезпечити комплексніший підхід, керований даними, гарантуючи, що вся необхідна інформація буде врахована під час обчислень.

Динаміка ринку нерухомості потребує здатності до адаптації в режимі реального часу, яку штучний інтелект вправно здійснює. Традиційні методи оцінки часто спираються на застарілі дані та не здатні враховувати ринкові зміни, які відбуваються швидко. Натомість штучний інтелект здатний плавно ураховувати зміни в реальному часі, що визначає його абсолютну перевагу [2].

Як наслідок, оцінка вартості, яку він виконує, ґрунтується не лише на минулих даних, але й відображає поточний стан ринку, допомагаючи зацікавленим сторонам краще зрозуміти середовище, яке постійно змінюється. Адаптація в режимі реального часу відрізняє штучний інтелект від традиційних підходів і підкреслює важливість його застосування на ринку, що стрімко розвивається, та є ознакою його революційної сили [3].

Переваги використання штучного інтелекту для визначення вартості нерухомості

Підвищений рівень точності. Використання штучного інтелекту відкриває потенціал уникнення суб'єктивності та упередженості, які довгий час супроводжували звичайні методології оцінювання власності. Це одна з найпереконливіших та найочевидніших переваг цього підходу. Алгоритми штучного інтелекту працюють у світі непідроблених фактів, вільних від впливу суб'єктивних думок чи емоційного впливу.

Така об'єктивність гарантує, що обчислення вартості ґрунтуватиметься виключно на самих даних, що дає змогу отримувати оцінку, яка надзвичайно точно відображає фактичну вартість майна на ринку. Це надає зацікавленим сторонам впевненість у чесності та справедливості угоди.

Ефективність у часі. У контексті оцінки майна час є надзвичайно цінним ресурсом. Традиційні методи потребують днів або навіть тижнів для ретельної оцінки, яку, безперечно, перевершують процеси, керовані штучним інтелектом, з погляду ефективності. Необхідний для оцінювання період значно зменшується завдяки неперевершеній швидкості, з якою системи штучного інтелекту аналізують дані. Такий підхід не лише прискорює транзакції, але й додає адаптивності процесу оцінки, щоб гравці швидко реагували на зміни ринкових умов.

Зниження витрат. Ефективність штучного інтелекту також впливає на фінансову складову оцінювання майна. Традиційні підходи часто потребують проведення кількох оцінок і навіть повторних переоцінок для підвищення точності, що може призвести до значних витрат. Натомість застосування методів машинного навчання зменшує потребу в повторних оцінках, ефективно знаходячи тенденції з різних джерел даних і синтезуючи їх. Як результат, це забезпечує істотне зниження витрат як для покупців, так і для продавців, а також перерозподіл коштів на важливіші процеси.

Використання геоданих. Геопросторові дані є ще одним важливим фактором оцінки власності, що охоплює все, починаючи від близькості об'єкта до зручностей і закінчуючи його розташуванням у зонах затоплення чи поблизу промислових зон. Алгоритми штучного інтелекту можуть інтегрувати геоінформаційні системи, щоб включити ці фактори в оцінки в реальному часі, пропонуючи рівень деталізації, якого раніше було важко досягти [4].

Чітке розуміння. Відкритість оцінок забезпечує прозорість в операціях з нерухомістю, якої не давало застосування традиційних методів. Алгоритми, які використовують в оцінці на основі штучного інтелекту, не є "чорними скриньками", оскільки вони пропонують стислі обґрунтування своїх висновків. Зацікавленим сторонам надається детальне пояснення кожної методології оцінювання. Така відкритість підвищує рівень довіри та сприяє прийняттю обґрунтованих рішень.

Покращений вибір інвестицій. Емпірична основа для оцінки на основі штучного інтелекту дає зацікавленим сторонам потужний інструмент для покращення інвестиційного вибору. Інвестори, покупці та продавці можуть використовувати інформацію, отриману за допомогою оцінки на основі даних, для прийняття зважених рішень. Точність оцінки, яку забезпечує штучний інтелект, допомагає зменшити ризики під час визначення очікуваної прибутковості нерухомості або аналізу ринкових тенденцій для здійснення стратегічного вибору. Такий спосіб прийняття рішень запобігає

можливостям переплатити за покупку або недооцінити вартість нерухомості у разі продажу, що забезпечує успішніші та обґрунтованіші результати [2].

Порівняння наявних систем з продажу, покупки та оренди нерухомості

Найпопулярнішими додатками з продажу, покупки та оренди нерухомості на українському ринку є DIM.RIA [5] та OLX [6].

Істотна перевага цих сервісів – їхня популярність в Україні, завдяки чому такі сервіси мають величезні обсяги даних про нерухомість, що постійно зростають. Цими даними можна користуватись, розробляючи такий сервіс, оскільки їх можна придбати за власні кошти.

На 10.09.2023 р. сервіс OLX містив такі активні оголошення: квартири у продажу – 190 447, будинки – 73 942, земельні ділянки – 35 147, комерційна нерухомість – 50 939, подобова оренда житла – 16 145. А DIM.RIA: квартири – 220 317, будинки – 31 110, земельні ділянки – 4 728, комерційна нерухомість – 31 548, загальна кількість нерухомості у продажу – 673 499, загальна кількість нерухомості для оренди – 117 873.

OLX слугує лише платформою для розміщення оголошень, не містить ніяких аналітичних інструментів, показників, послуг тощо. Охоплює надто велику кількість різноманітних сфер покупки та продажу.

Перевагою **DIM.RIA** є інтуїтивний інтерфейс, зручність пошуку і фільтрації оголошень, наявність функції сповіщень, орієнтованість на клієнта та налагодження взаємодії між експертами з продажу нерухомості та користувачами, відображення різноманітних показників мікрорайонів міста за десятибальною шкалою, зручностей та відстаней до них для конкретного об’єкта нерухомості. Є можливість безкоштовно визначити приблизну вартість нерухомості для авторизованого користувача, якщо він хоче її продати. В опис такої нерухомості входить велика кількість різних параметрів, що стосуються як самої нерухомості (площа, кількість кімнат, поверх, наявність балконів, якість ремонту, наявність утеплення), так і безліч додаткових факторів (наявність ліфту, місця паркування, дитячих майданчиків тощо). Невідомо, які з перелічених параметрів та характеристик насправді враховують під час ціноутворення. Видається, що сервіс лише повертає середнє значення вартості нерухомості зі схожими параметрами у певній області на карті, що не є достатньо точним показником для коректної оцінки нерухомості. Також корисною функцією для користувача є відображення вартості оренди та продажу наявної нерухомості за вибраними параметрами у вигляді графіка.

Найпопулярнішими додатками, використовуваними за кордоном, є Zillow, Realtor, Redfin, PropStream.

Zillow [7] поширений та актуальний переважно у Сполучених Штатах Америки і Канаді. Можна виділити такі його переваги: зручне відображення даних про одиницю нерухомості; наявність графіка зміни вартості нерухомості з часом (лише історія вартості); відображення корисних даних про район, в якому розташована нерухомість; відображається очікувана вартість здавання нерухомості в оренду. Недоліком є відсутність графіка з передбаченням зміни вартості нерухомості у майбутньому, що було б логічним продовженням відображення історичних даних.

Realtor [8] поширений та актуальний лише на території Сполучених Штатах Америки. Вирізняється з-поміж інших можливістю відображення різноманітних важливих даних на мапі. Основна перевага – відображення мапи, що показує розташування навчальних закладів з можливістю їх фільтрування за рейтингом, закладів харчування, руху громадського транспорту, велодоріжок, шумового забруднення та регіонів можливого затоплення у разі підвищення рівня води, що є дуже важливою інформацією для покупців нерухомості.

Redfin [9] поширений та актуальний тільки на території Сполучених Штатів Америки. Сервіс надає багато статистичної інформації у вигляді графіків, що може бути важливою під час прийняття рішення щодо покупки нерухомості. Можна виділити низку переваг цього додатка порівняно з конкурентами, а саме: відображення статистичних даних про нерухомість у вибраному районі міста за останні 30 днів; наявність графіка зміни середньої вартості нерухомості з часом; наявність

графіка кількості проданої нерухомості з часом; наявність графіка середньої кількості днів, потрібних для продажу нерухомості.

Сервіс **PropStream** [10] виділяється найбільшим функціоналом для аналізу даних. Додаток спеціально розроблений для інвесторів, брокерів, агентів з продажу нерухомості та інших експертів у цій сфері. Цей сервіс пропонує такі функції: відображення аналітичних звітів, пошук різноманітних документів стосовно власності, оцінювання вартості нерухомості, візуалізація перспективних районів на мапі, відображення історичного графіка зростання цін, визначення орендної вартості, відображення статистичних даних району, щомісячної зміни вартості та зміни орендної плати, оцінка доходу від оренди, оцінка власного капіталу, темпи зростання та багато інших. Переваги сервісу: спеціалізований додаток, що передбачає наявність багатьох аналітичних інструментів, корисних для інвесторів та різних експертів з операцій з нерухомістю; наявність інструментів email-сповіщень та розсилок; можливість виділяти на мапі район, в якому треба шукати чи аналізувати нерухомість; велика база даних з історією об'єктів нерухомості. Недоліки: складний функціонал, що потребує багато часу та підготовки, щоб розібратись у ньому та мати змогу повною мірою ним користуватись; застарілий користувацький інтерфейс; абсолютно усі функції платні; розрахований на вузьке коло користувачів.

Формулювання цілі статті

Мета статті – розробити алгоритми для надання користувачам актуальних даних щодо вартості об'єктів нерухомості. Це дасть змогу продавцям визначити поточну вартість нерухомості залежно від параметрів та розташування, з метою подальшого встановлення ціни продажу. Для покупців ця система буде актуальною, оскільки дасть змогу значно пришвидшити і спростити пошук нерухомості, надаючи інструменти фільтрування, оцінки, рекомендації, аналізу та автоматичного сповіщення.

Акцент на геопросторових даних, а отже, варто зазначити, що такі дані є одним з важливих факторів оцінки власності. Це охоплює все, починаючи від близькості об'єкта до зручностей і закінчуючи його розташуванням у зонах затоплення чи поблизу промислових зон. Алгоритми штучного інтелекту можуть інтегрувати геоінформаційні системи, щоб ввести ці фактори в оцінки в реальному часі, пропонуючи рівень деталізації, якого раніше було важко досягти [4].

Наукова новизна та інноваційність роботи полягає у використанні під час визначення вартості нерухомості моделей машинного навчання, а також у відкритості даних, що стосуються її вартості для широкого кола користувачів.

Основна відмінність застосованого підходу полягає у відсутності обмеження лише аналізом базових параметрів нерухомості, таких як площа, поверх чи кількість кімнат. Вона також враховує параметри розташування. Наприклад, відстань до центру міста, район міста, найважливішу інфраструктуру в радіусі 500 та 1000 метрів. Це дає змогу надавати точнішу й об'єктивнішу оцінку вартості нерухомості.

Крім того, реалізація моделі машинного навчання в інформаційній системі дасть змогу користувачам переглядати конкретні параметри, що найбільше вплинули на вартість нерухомості, а також саму оціночну вартість, яку обчислила система, що забезпечить прозорість та обґрунтованість такої оцінки.

Сьогодні на ринку немає аналогів цієї системи, враховуючи вказаний підхід для підрахунку вартості та надання такої детальної інформації користувачам.

Побудова моделі машинного навчання для розв'язання задачі прогнозування вартості нерухомості

Розроблювана модель буде використовуватися для оптимізації процесу пошуку нерухомості, оцінювання її вартості та надання можливості укладання взаємовигідних угод між покупцями та продавцями. Найважливіша мета проєкту полягає в застосуванні алгоритмів машинного навчання

для виконання функції оцінки вартості, що буде актуальним для усіх учасників ринку нерухомості. Важливо також досягнути якомога вищого рівня точності прогнозувань.

Складність завдання полягає в обмеженості кількості ресурсів, а особливо в закритості даних, що стосуються нерухомості. Недоліки українського ринку нерухомості в тому, що усі дані стосовно нерухомості є прихованими, зокрема історичні. Державна служба статистики України надає обмежену кількість інформації у цій галузі, яку неможливо використати у межах цієї роботи. Наприклад: індекс зростання цін на житлову нерухомість в Україні, порівняно з попереднім роком; середня вартість оренди однокімнатної квартири у різних областях України тощо.

У межах дослідження плануємо досягнути точності визначення вартості нерухомості в середньому 5–7 %. Значення похибки обчислюватимемо на основі тестових даних. Забезпечення такого рівня точності передбачень вважатиметься достатнім, оскільки витрати на послуги ріелтора в середньому становлять 3–5 % від загальної вартості продажу нерухомості [11, 12], а отже, їх майже повністю покриває похибка в розрахунках.

Процес створення моделі машинного навчання умовно поділено на кілька етапів, причому коректне та успішне виконання кожного попереднього етапу істотно впливає на наступний, як і на кінцевий результат загалом.

Перший етап

Спершу потрібно знайти всі доступні джерела даних, які можна використати в роботі. Таких даних має бути якомога більше, оскільки кількість є важливим критерієм для побудови моделі машинного навчання. Крім того, потрібно оцінити якість та повноту даних кожного джерела. В нашому випадку оголошення щодо продажу нерухомості повинно містити такі параметри, як загальна площа, площа кухні, поверх, загальна поверховість будинку, рік побудови будинку, адреса нерухомості тощо.

Знайшовши якісне джерело даних, потрібно ці дані завантажити локально, скориставшись офіційною API джерела. Такі послуги не є безкоштовними, а отже, потрібно також оплатити ключ, який дасть доступ до необхідної інформації.

Після цього потрібно виконати фільтрування даних: знайти і вилучити дублікати оголошень, а також визначити аутлаєри – об’єкти даних, які надто сильно виділяються з-поміж інших. Найпростішим аутлаєром є нерухомість, ціна за квадратний метр якої в кілька разів вища за переважну більшість записів, якщо у вибірці не існує ніяких інших даних зі схожими значеннями. Потім потрібно перевірити, чи є в оголошеннях пропущені дані та які. У разі, якщо немає не надто важливих даних, їх можна заповнити, в іншому випадку такі записи потрібно теж відфільтрувати.

Також для кожного об’єкта нерухомості потрібно виконати уніфікацію даних, тобто зведення їх до вигляду, що використовується в системі. Наступний крок – потрібно визначити координати нерухомості за адресою, використовуючи зовнішній API. Насамкінець потрібно доповнити географічні дані нерухомості, такі як кількість автобусних зупинок поруч, кількість лікарень, шкіл тощо.

Другий етап

Етап передбачає визначення мінімального переліку найважливіших ознак та параметрів нерухомості, що впливають на здатність моделі машинного навчання передбачати вартість нерухомості. Зазвичай цю процедуру виконують емпіричним способом, оскільки передбачити, як поводитиметься модель машинного навчання, оперуючи конкретним набором даних, практично неможливо.

Деякі окремі характеристики нерухомості потрібно видозмінити так, щоб їх ефективніше опрацьовувала модель. В цьому випадку доцільно застосувати метод унітарного кодування (англ. ***One-Hot Encoding***), що передбачає подання складних категоріальних ознак у вигляді простих значень 0 або 1. Наприклад, замість перелічення усіх типів стін будинку в одній колонці (цегла, газоблок, залізобетон, піноблок тощо) можна створити окрему колонку для кожного з типів стін, наприклад “цегляна_стіна”, значення якої буде 0 або 1. Якщо таких колонок буде багато, їх можна спробувати згрупувати за спільними ознаками. Це може позитивно вплинути на тренування моделі.

Третій етап

Потрібно розділити дані на дві вибірки – тренувальну і тестувальну, зазвичай це роблять у відсотковому співвідношенні 80/20 % відповідно. Бажано, щоб дані були хаотичними, тобто не посортованими та не згрупованими за жодним з параметрів. Особливо це стосується тих параметрів, що використовуються в процесі навчання моделі.

Для досягнення найкращих результатів можна спробувати вибирати дані для тестування методом перебору, перебираючи усі можливі варіанти. Такий спосіб називається перехресним затверджуванням (англ. *Cross-validation*). Потрібно брати 20 % даних для тестування із загальної сотні по черзі й тренувати модель тим самим алгоритмом на тих самих даних, проте вибираючи різні дані для тестування: 0–20 %, 20–40 %, 40–60 %, 60–80 %, 80–100 %. Так можна впевнитись, що для тренування вибрано найкращі дані, а також запобігти перенавчанню моделі.

До цього ж етапу можна зарахувати і сам процес тренування моделей машинного навчання. Оскільки активних оголошень для одного великого міста в межах від декількох тисяч до кількох десятків тисяч (до фільтрування), таку кількість вважають порівняно невеликою і очікуване тренування моделі повинно тривати секунди або хвилини. Отже, оптимальним рішенням буде не обмежуватись одним конкретним алгоритмом, а одразу спробувати декілька і перевірити, який з них працюватиме ефективніше. На цьому етапі алгоритми слід запускати з параметрами, що встановлені за замовчуванням.

Четвертий етап

Потрібно оцінити точність роботи моделей, використовуючи одну з метрик якості. Дізнавшись, які алгоритми найкраще працюють для вирішення цього завдання, необхідно вибрати один або кілька з них і спробувати підібрати оптимальні гіперпараметри. Так можна досягнути високої точності роботи моделі для конкретного завдання.

Очевидно, що алгоритми регресії машинного навчання за такого підходу беруть з готових бібліотек, оскільки вказана послідовність кроків передбачає одночасну роботу з кількома з них.

Прийнято рішення обмежити кількість міст для обчислення вартості нерухомості до одного.

Вибір та обґрунтування методів розв’язання задачі

Оскільки завдання полягає у визначенні вартості нерухомості на підставі певних вхідних параметрів, задачу такого типу допомагають розв’язати алгоритми регресії [15–28].

Регресія – це підхід, що передбачає дослідження зв’язку між незалежними змінними або ознаками та залежною змінною або результатом. Регресію використовують у машинному навчанні як метод предиктивного моделювання, тобто для прогнозування чи передбачення результатів.

У машинному навчанні є різні підходи, які використовують для виконання регресії, також різні популярні алгоритми. Різні методики можуть використовувати різну кількість незалежних змінних або обробляти різні типи даних. Різні типи регресійних моделей машинного навчання можуть передбачати різне співвідношення між незалежними та залежними змінними.

Проста лінійна регресія (*Linear Regression*) – це статистичний метод встановлення зв’язку між двома змінними за допомогою прямої лінії. Лінію рисують, знаходячи нахил та перетин, що визначає лінію та мінімізує помилки регресії [15–19].

Множинна лінійна регресія (*MLR*). Під час прогнозування результату складного процесу найкраще використовувати множинну лінійну регресію замість простої. Для виконання простих завдань не обов’язково використовувати складні алгоритми.

Проста лінійна регресія може точно охопити зв’язок між двома змінними у простих співвідношеннях. Але коли ви маєте справу зі складними завданнями, де декілька незалежних параметрів впливають на результат, тоді потрібно перейти від простої до множинної регресії.

Модель множинної регресії використовує формулу, що містить більше ніж одну незалежну змінну, тому здатна працювати з кривими, а також з нелінійними залежностями [20].

Модель у лінійній регресії будують, мінімізуючи суму квадратичних відхилень між спостережуваними та розрахунковими значеннями (методом найменших квадратів).

Дерево рішень (*Decision tree*). У машинному навчанні використовують як модель прогнозування (регресійний аналіз дерева), що відображає знання про об'єкт (подані "гілками" (ребра)) у множину рішень. Листя дерева – значення цільової функції, а вузол – умови переходу, що визначають, по якій з гілок рухатись. Якщо для цього спостереження умова набуває істинного значення, то перехід здійснюється по лівому ребру, якщо ж хибного – то по правому.

Дерево рішень розростається у результаті ітераційного розбиття його вузлів доти, доки "листя" не міститиме більше розгалужень (відповідь на останнє питання дає однозначний результату), або поки не буде досягнуто якоїсь умови завершення. Створення дерева рішень починається з кореня дерева, дані розподіляють так, щоб отримати найбільше значення інформаційного приросту (англ. *Information Gain, IG*) [19]. Набір даних є чистим або однорідним, якщо містить лише один клас (ТАК або НІ). Якщо набір даних містить кілька класів – таблиця нечиста або неоднорідна (комбінація ТАК і НІ) [23, 24].

"Випадковий ліс" (*Random Forest*). Якщо об'єднати кілька некорельованих дерев рішень, то часто можна досягти істотного підвищення точності моделі. Такий метод називають "випадковим лісом". Під час росту на розгалуження дерева впливають певні випадкові процеси (це називають рандомізацією). Остаточна модель відображає усереднення дерев.

За словами Бреймана, який ввів термін "випадковий ліс" у 1999 р., є різні методи рандомізації. Створюють цей ліс так. Спочатку вибирають випадковий екземпляр із загального набору даних для кожного дерева. У міру зростання дерева відбувається вибір підмножини особливих властивостей (англ. *features*) у кожному вузлі. Це є критерієм поділу набору даних. Після цього окремо для кожного дерева рішень визначають цільове значення. Усереднення значень цих передбачень надає кінцеве передбачення алгоритму "випадкового лісу" [19].

Оскільки цей алгоритм об'єднує кілька моделей в одну, він належить до сфери "ансамблевого навчання" (*Ensemble Learning*). А якщо точніше, *Random Forest* – це так звана техніка бегінг (*Bagging*).

Метод опорних векторів (*Support Vector Regression*). Функціональність методу опорних векторів (SVR) ґрунтується на опорній векторній машині (англ. *Support Vector Machine*, скорочено SVM). Розглянемо простий приклад. Потрібно знайти лінійну функцію:

$$f(x) = \langle \mathrm{w}, \mathrm{x} \rangle + \mathrm{b}\,, \tag{1}$$

де ⟨w, x⟩ описує перехресний добуток.

Мета SVR – знайти пряму лінію як модель для точок даних, тоді як параметри лінії мають бути визначені так, щоб лінія була настільки "плоскою", наскільки можливо. Цього можна досягти мінімізацією норми:

$$\|w\|_2 \coloneqq \sqrt{(w_1)^2 + (w_2)^2 + \cdots + (w_n)^2} = \left(\sum_{i=1}^{n} (w_i)^2\right)^{1/2}. \tag{2}$$

Під час побудови моделі не має значення, наскільки далеко точки даних від змодельованої прямої, якщо вони містяться в межах визначеного діапазону (від –e до +e). Відхилення, що перевищують задану межу e, не допускаються [19].

Метод найближчого сусіда (*K-Nearest Neighbors* or *KNN*). KNN – це непараметричний, простий, але потужний алгоритм навчання з учителем, який можна використовувати як для завдань регресії, так і для класифікації. Основна ідея KNN полягає у тому, щоб знайти K найближчих точок даних у навчальному просторі для нової точки даних. Зробивши це, можна зарахувати нову точку даних до одного з класів, проаналізувавши, який клас переважає серед *k* найближчих сусідніх точок [31]. У випадку регресії залежна змінна неперервна, вона розсіяна по всій координатній площині. Коли є нова точка даних, кількість сусідів (K) визначають за допомогою будь-якої метрики відстані. Після знаходження сусідів прогнозоване значення нової точки даних є середнім значенням усіх сумісних сусідів.

Наприклад, розглянемо прогноз ціни на будинок. Ціна – це залежна змінна, а площа в квадратних метрах будинку – незалежна змінна. Тепер, після відображення всіх точок даних на

декартовій площині, коли з'являється нова точка у квадратних метрах, середня вартість сусідів K за квадратний метр будинку є вартістю нової точки даних. Тому замість прогнозування класу регресор використовує середнє значення всіх сусідніх точок [31].

Вибір значення K. Значення *K* є основною частиною KNN і його вибір може бути складним. Послідовність пошуку найкращого та оптимального значення *K*:

- потрібно розділити набір даних на навчальну та тестувальну вибірки;
- вибрати діапазон *K* значень. Можна почати з *K* = 1 і поступово його збільшувати;
- розпочати навчання моделі KNN для кожного значення *K*;
- оцінити продуктивність навчених моделей, використовуючи діапазон значень *K*.

Важливо зауважити, що вибір значення *K* залежить від набору даних і самої задачі. Мале значення *K* може призвести до того, що модель не буде достатньо гнучкою, що характеризує поняття перенавчання (*overfitting*), тоді як велике значення може призвести до недонавчання (*underfitting*). Тому рекомендують експериментувати з різними значеннями *k*, щоб знайти оптимальне із них для конкретного набору даних.

Переваги алгоритму:

- Гнучкість: застосовують як для розв'язання задач регресії, так і для класифікації.
- Відсутність потреби навчання: економить час та обчислювальні ресурси.
- Стійкість до шумних даних, оскільки він покладається на більшість голосів найближчих сусідів. Це робить його менш вразливим до аутлаєрів.
- Добре працює з невеликими даними: для прогнозування не потрібна велика кількість даних. Непараметричність: KNN є непараметричним алгоритмом, тобто не потрібно жодних припущень щодо даних.
- Простота: KNN – простий і зрозумілий алгоритм.

Недоліки:

- Оптимальне значення K. Вибір правильного значення K важливий, оскільки це вплине на продуктивність моделі. Немає універсального значення K, і це значення залежить від характеристик елементів набору даних.
- Незбалансовані дані:у KNN можливі відхилення у випадку незбалансованих даних. Тобто коли один клас повинен мати більше прикладів, ніж інший, KNN може передбачити мажоритарний клас для тестових прикладів.

Проблема розмірності: KNN може страждати від проблеми високої розмірності, яка виникає, коли кількість функцій велика. Зі зростанням кількості вимірів відстань між будь-якими двома точками в даних має тенденцію до збільшення, що ускладнює пошук значущих найближчих сусідів [31].

Метрики якості регресійних моделей. Для того, щоб модель лінійної регресії можна було застосовувати на практиці, спершу необхідно оцінити її якість. Для цього є певні показники, кожен з яких призначений для використання у різних ситуаціях і має певні особливості застосування (лінійні та нелінійні регресії, стійкі до аномалій, абсолютні та відносні тощо). Коректний вибір міри з метою оцінки якості моделі – це один із найважливіших факторів успіху під час вирішення завдань аналізу даних.

"Хороша" аналітична модель повинна задовольняти дві вимоги, що часто суперечать одна одній – якнайкраще відповідати даним і бути зручною для інтерпретації користувачем. Підвищення відповідності моделі даним зазвичай пов'язане з її ускладненням, оскільки у випадку регресії це означає збільшенням кількості вхідних змінних моделі. А що складніша модель, то складніше її інтерпретувати.

Тому в разі вибору між простою та складною моделлю остання повинна значуще збільшувати відповідність моделі даним, щоб виправдати зростання складності та відповідне зниження інтерпретованості. Якщо ця умова не виконується, потрібно вибрати простішу модель [32].

Отже, щоб оцінити, наскільки підвищення складності моделі збільшує її точність, необхідно використати інструмент оцінки якості регресійних моделей, що містить такі міри: середньоквадратична помилка (MSE); Корінь із середньоквадратичної помилки (RMSE); середньоквадратична помилка у відсотках (MSPE); середня абсолютна помилка (MAE); середня абсолютна помилка у відсотках (MAPE); симетрична середня абсолютна відсоткова помилка (SMAPE); середня абсолютна масштабована помилка (MASE); середня відносна помилка (MRE); середньоквадратична логарифмічна помилка (RMSLE); коефіцієнт детермінації R-квадрат; коригований коефіцієнт детермінації.

Розроблення алгоритмів розв'язання задачі

Деякі дані з вибірки потребують заповнення пропущених значень, для того щоб уникнути їх фільтрації. Це можна зробити, використовуючи значення середнього, моди та медіани загальної вибірки [33]. Вибір оптимального підходу усереднення значень залежить від виду даних, а також від їх розподілу у вибірці.

Використано лише одну метрику оцінки якості роботи моделі машинного навчання, а саме – R-квадрат. Цю метрику також називають коефіцієнтом детермінації, що показує частку дисперсії залежної змінної, яку пояснюють за допомогою регресійної моделі. Особливість цього підходу полягає у тому, що він показує, наскільки ця модель працює краще, ніж модель, в якій є тільки константа, а вхідні змінні відсутні або коефіцієнти регресії при них дорівнюють нулю.

Найзагальніша формула для обчислення коефіцієнта детермінації така [34]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n}(\overline{y_i} - y_i)^2}\ , \tag{3}$$

де y_i – фактичне значення; $\widehat{y_i}$ – значення, обчислене моделлю; $\overline{y_i}$ – розраховане за формулою середнього арифметичного:

$$\overline{y_i} = \frac{1}{n}\sum_{i=1}^{n} y_i\ . \tag{4}$$

Ґрунтуючись на формулі (3), в контексті створення регресійної моделі машинного навчання, можна дати спрощене визначення коефіцієнта детермінації – це метрика, що показує, наскільки значення, які прогнозує регресор, є "кращими", ніж якби модель просто розраховувала середнє значення уо всій вибірці.

Щоб оцінити результат, який повертає формула R-квадрат, на практиці використовують таку шкалу [34]:

- значення, менші за 0,5 (ураховуючи від'ємні числа), вказують на те, що модель погана;
- значення >0,5 свідчать про задовільність моделі;
- якщо коефіцієнт детермінації >0,8, то модель вважають хорошою.

З формули зрозуміло, що коли передбачені значення $\widehat{y_i}$ дорівнюватимуть середньому арифметичному $\overline{y_i}$, то результат обчислень у дробі дорівнюватиме 1, а отже, коефіцієнт детермінації становитиме 0, що вказує на погані результати роботи моделі. Якщо $\widehat{y_i}$ будуть дорівнювати y_i, то це дасть 0 у чисельнику і коефіцієнт детермінації дорівнюватиме 1, а це означатиме, що модель ідеально передбачає усі значення [34].

Тестування моделі машинного навчання

Тестування є важливим, оскільки надає можливість виявити і виправити помилки у системі на етапі її розроблення, а також допомагає зрозуміти, наскільки система готова до використання.

Пошук, фільтрування та оброблення даних. Першим і надзвичайно важливим кроком, який визначає усі подальші можливості системи, є пошук актуальних даних ринку нерухомості. Кількість і якість даних, доступних для навчання та тестування моделі машинного навчання, відіграють істотну роль в її ефективності. Дані можуть мати різні форми, такі як числові, категоріальні (які можна використати для класифікації об'єктів на основі їх належності до категорій) або часові ряди

даних (послідовність значень, зібраних відповідно до певного часового інтервалу), і можуть надходити із різних джерел, таких як бази даних, електронні таблиці або API [35].

Потрібно знайти якомога більше джерел інформації, що забезпечать наявність достатньої кількості даних для формування якісної моделі машинного навчання. Також бажано оновлювати ці дані з певною періодичністю, для того щоб мати надалі можливості їх глибшого аналізу, забезпечення функції сповіщення і загалом перетренувати модель на актуальних даних.

Прийнято рішення зібрати усі необхідні дані лише для одного конкретного міста, оскільки збирання даних є дуже затратним, особливо через те, що для кожного міста потрібно тренувати окрему модель машинного навчання для забезпечення точнішої роботи програми. Це зумовлено тим, що ціна на нерухомість помітно відрізняється в різних містах України. Наприклад, середня вартість однокімнатної квартири у Львівській області – 61 616 доларів США, тоді як у Миколаївській – 23 144 дол., відповідно до відкритих джерел інформації на серпень 2023 р. [36]. Тому, якщо спробувати створити одну універсальну модель для всієї України, вона не зможе виконати адекватну оцінку вартості за вхідними даними, а також це призведе до непередбачуваних результатів для аутлаєрів моделі.

Застосовано декілька різних підходів до збирання даних. Перший з них – видобування даних із офіційних API вебсайтів: DIM.RIA [7] та OLX [8]. Також було знайдено такі телеграм-канали, що публікують оголошення з продажу та оренди нерухомості: Нерухомість Київ та область [37], Нерухомість Київської області [38].

За допомогою відповідної бібліотеки Telethon [39] ці дані вдалось зібрати та обробити для подальшого використання. Наведемо фрагмент коду, що аналізує текст оголошення оренди нерухомості та видобуває дані з нього:

```
message_text = message_text.replace('\n', ' ')
address_pattern = r'\[.+\]'
square_pattern = r'[0-9 ]+м²'
ad_id_pattern = r'№[0-9a-zA-Z ]+'
price_pattern = r'Ціна[0-9a-zA-Z \t]+'
floor_pattern = r'[0-9й \t]+поверх'
rooms_pattern = r'#[0-9]кімнатна'
address = re.search(address_pattern, message_text).group(0) if re.search(address_pattern, message_text) else 'None'
square = re.search(square_pattern, message_text).group(0) if re.search(square_pattern, message_text) else 'None'
square = square.replace('м²', '').strip()
ad_id = re.search(ad_id_pattern, message_text).group(0) if re.search(ad_id_pattern, message_text) else 'None'
price = re.search(price_pattern, message_text).group(0) if re.search(price_pattern, message_text) else 'None'
floor = re.search(floor_pattern, message_text).group(0) if re.search(floor_pattern, message_text) else 'None'
rooms = re.search(rooms_pattern, message_text).group(0) if re.search(rooms_pattern, message_text) else 'None'
return ad_id, square, price, address, floor, rooms, message_text
```

Необхідно також виконати їх фільтрацію. Деякі дані не мають практичної цінності, оскільки містять недостатню кількість інформації в описі про одиницю нерухомості. Тому, якщо одне з вказаних вище полів не заповнено, такі дані не враховуються. Проте є винятки, за яких дані з пропущеними значеннями можна використати для побудови моделі: якщо в одиниці нерухомості немає значення площі кухні або житлової площі, то такі дані заповнюють медіаною цих значень у загальній вибірці. Середнє арифметичне значення брати недоцільно, оскільки розподіл цих даних – косий. Крім того, було заповнено значення типу опалення, типу оголошення і типу стін модою даних у вибірці, оскільки це категоріальні типи даних. Також за допомогою моди було заповнено і рік побудови будинку:

```
df.loc[df['kitchen_square']!= -1, 'kitchen_square'].median()
df.loc[(df['living_square']!= -1), 'living_square'].median()
df.loc[(df['wall_type']!= 'None'), 'wall_type'].mode()
df.loc[(df['building_year']!= -1), 'building_year'].mode()
df.loc[(df['advertisement_type']!= -1), 'advertisement_type'].mode()
df.loc[(df['heating_type']!= -1), 'heating_type'].mode()
```

Було визначено такі дані для доповнення інформації про нерухомість у місті Київ: площа кухні – 13 метрів квадратних, житлова площа – 30, тип стін – цегла червона, рік побудови будинку – 2013, тип оголошення – від посередника, тип опалення – централізоване.

Для того, щоб перевести адресу нерухомості в координати на мапі, ми використали відкрите джерело OpenStreetMap Nominatim API [40]. Ним можна користуватись без реєстрацій, потрібно лише надіслати GET запит та додати адресу в посилання:

```
https://nominatim.openstreetmap.org/search?q=Хрещатик+Київ+18&format=geojson
```

Так було визначено, що координати адреси вулиці Хрещатик, 18, у місті Київ такі: широта – 50.4510346, довгота – 30.523997621368913.

Потрібно знайти усі важливі локації в заданому радіусі поблизу об'єкта нерухомості. Для цього було використано відкрите джерело даних – Python Overpass API [41]. Щоб використовувати цей API, потрібно створити відповідний запит, що містить координати конкретної точки, а також радіус, у межах якого варто шукати найближчі локації. Нижче наведено приклад пошуку локацій на мапі у радіусі 100 метрів до заданої точки:

```
lat = 50.450310
lon = 30.523736
query = """
  (node(around:100,{lat},{lon});
  );out;
  """.format(lat=lat,lon=lon)
api = overpy.Overpass()
result = api.query(query)
file_path = "overpass_results.txt"
with open(file_path, "w", encoding="utf-8") as file:
  filtered_nodes = [node for node in result.nodes if node.tags]
  for node in filtered_nodes:
    file.write(f"Node {node.id} at ({node.lat}), {node.lon}, tags: {node.tags})\n")
```

Отримано 127 локацій у радіусі 100 метрів до заданої адреси. Очевидно, що ці дані потрібно буде відфільтрувати, оскільки в список локацій ввійшли дерева, сміттєві баки, вуличні лампи, поштові скриньки, лавочки тощо. Проте серед них є і важливі локації, які повинні впливати на вартість нерухомості.

Тепер, коли дані зібрано, відфільтровано і доповнено додатковими параметрами, можна розпочати **тренування моделі**. Ми вирішили використовувати відразу кілька алгоритмів регресії машинного навчання, щоб знайти той, що розв'яже таку задачу.

Алгоритми було запущено зі стандартними параметрами й отримано такий результат (рис. 1):

```
LinearRegression            0.35640526160283725
DecisionTreeRegressor       0.3287186211464418
KNeighborsRegressor         0.2439190832334115
SupportVectorRegression     0.31189351581228564
GradientBoostingRegressor   0.6345036822381225
RandomForestRegressor       0.6815511209834788
```

Рис. 1. Порівняння точності моделей, натренованих різними алгоритмами, за допомогою метрики R-квадрат

Порівнюючи коефіцієнти детермінації отриманих результатів, можна зробити висновок, що алгоритм "випадковий ліс" дав найкращий результат. Тому надалі розроблення моделі та підбір гіперпараметрів буде виконано саме для цього алгоритму.

Емпіричним способом підібрано такі гіперпараметри, що дали покращені результати передбачення прогнозованих значень, використовуючи алгоритм "випадковий ліс":

```
predictor_rf = RandomForestRegressor(n_estimators=200,
                                     min_samples_leaf=2,
                                     max_depth=6)
```

За рахунок конфігурування гіперпараметрів, а також використання методу перехресного затверджування даних (*Cross-validation*) значення R-квадрат вдалось підвищити до 0.81.

Процес **тестування** важливий, оскільки надає можливість виявити та виправити помилки у системі на етапі її розроблення, а також допомагає зрозуміти, наскільки система готова до використання. Створено валідаційну вибірку даних про нерухомість та перевірено, як модель справляється із прогнозуванням вартості, порівняно з реальними даними.

Рис. 2. Результат прогнозування вартості метра квадратного нерухомості залежно від її параметрів у місті Київ

Загалом модель вловлює кореляції між характеристиками нерухомості та її ціною (рис. 2). Проте таке подання результатів не є інформативним і не дає змоги повною мірою оцінити можливості моделі.

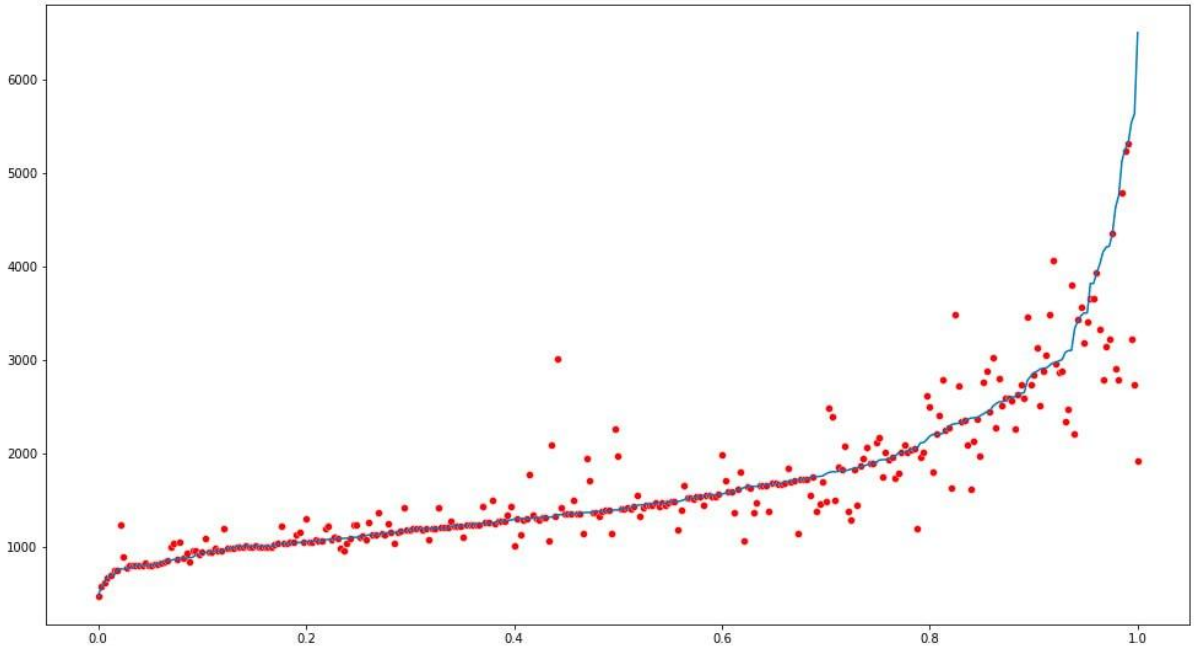

Рис. 3. Результат прогнозування вартості метра квадратного нерухомості за її параметрами у Києві розробленою моделлю

На рис. 3 вертикальна вісь відображає вартість нерухомості за квадратний метр у доларах США, а горизонтальна – пропорцію кількості оголошень. Червоними точками відображено значення вартості, передбаченні моделлю, а синіми – реальні значення.

На графіку спостерігається явище, що називається *гетероскедастичністю*, яке означає, що зі зростанням залежної величини прогнозована змінна матиме більші відхилення від реальних значень. Тобто зі зростанням вартості нерухомості за квадратний метр збільшується варіативність прогнозованих значень. Ця проблема пов'язана з недостатньою кількістю даних, використаних для навчання моделі, й це нормальне явище. Вартість за квадратний метр більшості нерухомості у вибірці даних менша від 2000 доларів, тому моделі складно передбачити такі унікальні випадки. В майбутньому цю проблему можна буде вирішити, збираючи все більше даних про нерухомість у продажу.

Для того, щоб краще оцінити результати обчислення абсолютної похибки, дані наведено у такому вигляді:

APE > 50 % for 0.012084592145015106 of test data
APE > 20 % for 0.1933534743202417 of test data
APE > 10 % for 0.31722054380664655 of test data
APE > 5 % for 0.4108761329305136 of test data
APE > 1 % for 0.5347432024169184 of test data
APE < 1 % for 0.4652567975830816 of test data,

де APE – абсолютна похибка, яку обчислюють за формулою:

$$APE = \left|\frac{d_i - \widehat{d_i}}{d_i}\right|, \quad (5)$$

де d_i – реальне значення; $\widehat{d_i}$ – значення, обчислене моделлю.

Тепер відсоткове співвідношення абсолютної похибки прогнозування вартості нерухомості можна подати у вигляді діаграми (рис. 4).

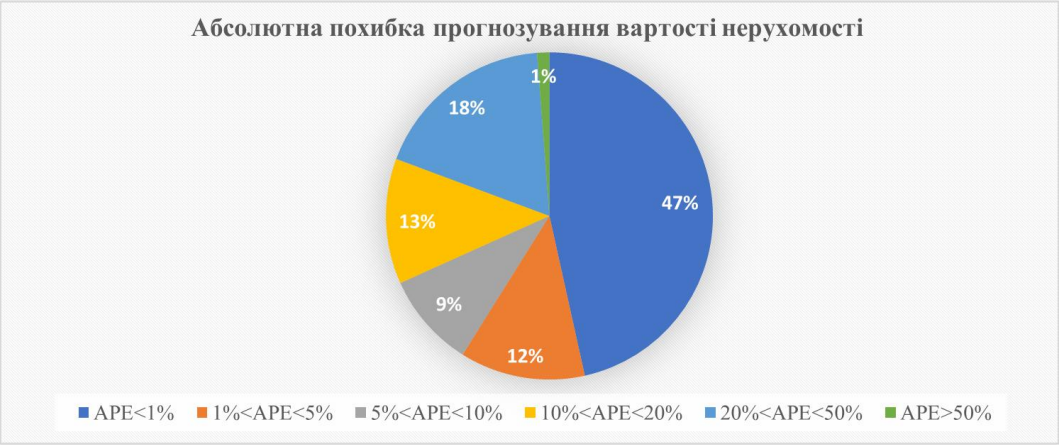

Рис. 4. Абсолютна похибка прогнозованої вартості нерухомості

На рис. 4 зображено розподіл абсолютної похибки прогнозованих результатів у відсотках. Середнє значення похибки становить 8,49 %, а медіана дорівнює 1,9 %.

Підсумовуючи все, що стосується моделі машинного навчання, можна зазначити, що загалом модель працювала добре, результати дослідження свідчать про те, що випробування пройшло успішно. За умови постійного зберігання нових даних, а також перетренування моделі на них можна сподіватись навіть на кращі результати її роботи.

Висновки

Результатом дослідження є створена модель машинного навчання, за допомогою якої можна визначити вартість нерухомості, залежно від її фізичних параметрів та географічного розташування. Проаналізовано основні переваги застосування методів штучного інтелекту для прогнозування вартості нерухомості. Проаналізовано українські та закордонні застосунки для вирішення проблеми

прогнозування вартості та оцінювання нерухомості. Особливість розробки полягає у тому, що для тренування моделі використовують не лише базові характеристики нерухомості, але й геопросторові дані. Враховується наявність та кількість локацій певного типу в заданому радіусі від об'єкта нерухомості, що дає змогу із вищою точністю прогнозувати вартість.

Створення моделі машинного навчання умовно розкладено на чотири етапи, а саме: збирання даних, їх фільтрування, обробку, доповнення, розподіл на різні вибірки і тренування моделі на основі цих даних. Досліджено методи та алгоритми регресії, а також метрики якості регресійних моделей з погляду задачі прогнозування вартості нерухомості. Перевірено якість роботи моделі машинного навчання на валідаційній вибірці даних, результати прогнозів і абсолютну похибку візуалізовано за допомогою графіків та діаграм. Проаналізувавши цю інформацію, ми зробили висновок, що результати роботи моделі відповідають вимогам системи, а отже, дослідження пройшло успішно.

Така інформаційна система спростить та пришвидшить пошук оптимальної нерухомості для покупки та оцінки її вартості для подальшого продажу. Впровадження системи в сферу нерухомості буде корисним для покупців та продавців нерухомості, а також для агентів і інвесторів, спрощуючи їхні бізнес-процеси.

Можна виділити такі позитивні ефекти, які дасть реалізація системи:

- *Спрощення пошуку нерухомості*: система дасть змогу покупцям знайти нерухомість, яка відповідає їхнім потребам та можливостям, швидше і ефективніше.
- *Покращення оцінки вартості*: продавці нерухомості зможуть точніше оцінити вартість своєї власності, що допоможе їм знайти покупців та укласти вигідні угоди.
- *Зниження ризику для інвесторів*: інвестори зможуть користуватися системою для аналізу та вибору об'єктів нерухомості із вищим потенціалом.
- *Позитивний вплив на ринок нерухомості*: усі зацікавлені сторони (покупці, продавці, агенти та інвестори) отримають інструменти, які сприятимуть підвищенню ефективності ринку нерухомості та полегшать взаємодію між ними.

Подальші дослідження будуть спрямовані на реалізацію та тестування прототипу інформаційної системи. Прийнято рішення розділити систему на три окремі сервіси, кожен із яких відповідатиме за певний перелік функцій. Розгортання системи заплановано виконати із контейнеризацією та налаштуванням створених контейнерів, використовуючи можливості Docker.

Список літератури

1. (n. d.). *The Importance of Accurate Property Valuation in Real Estate*. Sugermint.com. https://sugermint.com/the-importance-of-accurate-property-valuation-in-real-estate/

2. (n. d.). *AI in real estate property valuation: Is it really a game-changer?* Mdevelopers.com. https://mdevelopers.com/blog/ai-real-estate-property-valuation

3. Kolesnikova, I. (2023). *Using Artificial Intelligence for Real Estate: A Comprehensive Guide*. Mindtitan.com. https://mindtitan.com/resources/industry-use-cases/artificial-intelligence-in-real-estate/

4. (2023, September 22). *Real-time property valuations: how ai algorithms are making it possible*. Realspace3d.com. https://www.realspace3d.com/blog/real-time-property-valuations-how-ai-algorithms-are-making-it-possible/

5. Veres, O., Ilchuk, P., & Kots, O. (2021). *Data Science Methods in Project Financing Involvement*, In 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT) (Vol. 2, pp. 411–414). DOI: 10.1109/CSIT52700.2021.9648679

6. Veres, O., Ilchuk, P., & Kots, O. (2023). *Data Analytics on Debt Financing Research Based on Scopus and WoS Metrics*, In 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT). DOI: 10.1109/CSIT61576.2023.10324179

7. (n. d.). *DIM.RIA – вся нерухомість України. Продаж та оренда будь-якої нерухомості*. Dom.ria.com. https://dom.ria.com/uk/

8. (n. d.). *Український сервіс оголошень*. Olx.ua. https://www.olx.ua/uk/nedvizhimost/

9. (n. d.). *Agents. Tours. Loans. Homes.* Zillow. https://www.zillow.com/

10. (n. d.). *For Sale, Real Estate & Property Listing*. Realtor.com. https://www.realtor.com/

11. (n. d.). *Real Estate, Homes for Sale, MLS Listings, Agents*. Redfin.com. https://www.redfin.com/

12. (n. d.). *Most Trusted Provider of Real Estate Information*. Propstream.com. https://www.propstream.com/

13. Бережна, Н. (2021). *Купівля житла: чи потрібен рієлтор і скільки коштують його послуги в Україні*. Realestate.24tv.ua. https://realestate.24tv.ua/kupivlya-zhitla-potriben-rieltor-skilki-koshtuyut-ostanni-novini_n1525065

14. (n. d.). *How much is my home worth?* Zillow.com. https://www.zillow.com/how-much-is-my-home-worth/

15. (n. d.) *Machine Learning Regression Explained*. Seldon.io. https://www.seldon.io/machine-learning-regression-explained

16. Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

17. Finch, W. H., Bolin, J. E., & Kelley, K. (2019). *Multilevel modeling using R*. Crc Press

18. Evans, C., Leckie, G., & Merlo, J. (2020). Multilevel versus SingleLevel Regression for the Analysis of Multilevel Information: The Case of Quantitative Intersectional Analysis. *Social Science and Medicine*, 245, 112499. Article 112499. https://doi.org/10.1016/j.socscimed.2019.112499

19. (n. d.). *What is Simple Linear Regression in Machine Learning?* Simplilearn.com. https://www.simplilearn.com/what-is-simple-linear-regression-in-machine-learning-article

20. Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147.

21. Polzer, D. (2021). *7 of the Most Used Regression Algorithms and How to Choose the Right One. Linear and Polynomial Regression, RANSAC, Decision Tree, Random Forest, Gaussian Process and Support Vector Regression*. Towardsdatascience.com. https://towardsdatascience.com/7-of-the-most-commonly-used-regression-algorithms-and-how-to-choose-the-right-one-fc3c8890f9e3

22. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

23. Dawson, C. (2021, January 23). *Understanding Multiple Linear Regression*. Medium.com. https://medium.com/swlh/understanding-multiple-linear-regression-e0a93327e960

24. Mahaboob, B., Praveen, J. P., Rao, B. A., Haranadh, Y., Narayana, C., & Prakash, G. B. (2020). A study on multiple linear regression using matrix calculus. *Advancecs in Mathematics Scientifc journal*, *9*(7), 1–10.

25. Bouzebda, S., Souddi, Y., & Madani, F. (2024). Weak Convergence of the Conditional Set-Indexed Empirical Process for Missing at Random Functional Ergodic Data. *Mathematics*, *12*(3), 448.

26. Zhou, Y., & He, D. (2024). Multi-Target Feature Selection with Adaptive Graph Learning and Target Correlations. *Mathematics*, *12*(3), 372.

27. Li, T., Frank, K. A., & Chen, M. (2024). A Conceptual Framework for Quantifying the Robustness of a Regression-Based Causal Inference in Observational Study. *Mathematics*, *12*(3), 388.

28. Leyland, A. H., & Groenewegen, P. P. (2020). *Multilevel Modelling for Public Health and Health Services Research: Health in Context* (p. 286). Springer Nature. https://doi.org/10.1007/978-3-030-34801-4.

29. (n.d.) *Measure of impurity*. Medium.com. https://medium.com/@viswatejaster/measure-of-impurity-62bda86d8760

30. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967.

31. Rishit, P. (2023). *Understanding K-Nearest Neighbors: A Simple Approach to Classification and Regression. Demystifying K-Nearest Neighbors: Unveiling the Power of Proximity-based Algorithm*. Pub.Towardsai.net. https://pub.towardsai.net/understanding-k-nearest-neighbors-a-simple-approach-to-classification-and-regression-e4b30b37f151

32. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

33. Bhandari, P. (2020). *Central Tendency | Understanding the Mean, Median & Mode*. Scribbr.com. https://www.scribbr.com/statistics/central-tendency/

34. (n. d.). *What Is R Squared And Negative R Squared*. Fairlynerdy.com. http://www.fairlynerdy.com/what-is-r-squared/

35. (n. d.) *ML | Introduction to Data in Machine Learning*. Geeksforgeeks.org. https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/

36. (n. d.)). *Які середні ціни на вторинному ринку житла по Україні: скільки доведеться віддати за однокімнатну квартиру*. Sud.ua. https://sud.ua/uk/news/ukraine/259841-kakie-srednie-tseny-na-vtorichnom-rynke-zhilya-po-ukraine-skolko-pridetsya-otdat-za-odnokomnatnuyu-kvartiru

37. (n. d.). *Нерухомість Київ та область*. https://t.me/ppbestate

38. (n. d.). *Нерухомість Київської області*. https://t.me/Neruhomist_Kyiv_region

39. (n. d.). *Telethon's Documentation*. Docs.Telethon.dev. https://docs.telethon.dev/en/stable/

40. (n. d.). *Nominatim 4.3.0 Manual*. Nominatim.org. https://nominatim.org/release-docs/latest/api/Overview/

41. (n. d.). *Welcome to Python Overpass API's documentation!* Python-Overpy.Readthedocs.io. https://python-overpy.readthedocs.io/en/latest/

References

1. (n. d.). *The Importance of Accurate Property Valuation in Real Estate*. Sugermint.com. https://sugermint.com/the-importance-of-accurate-property-valuation-in-real-estate/

2. (n. d.). *AI in real estate property valuation: Is it really a game-changer?* Mdevelopers.com. https://mdevelopers.com/blog/ai-real-estate-property-valuation

3. Kolesnikova, I. (2023, March 31). *Using Artificial Intelligence for Real Estate: A Comprehensive Guide.* Mindtitan.com. https://mindtitan.com/resources/industry-use-cases/artificial-intelligence-in-real-estate/

4. (n. d.) *Real-time property valuations: how ai algorithms are making it possible*. Realspace3d.com. https://www.realspace3d.com/blog/real-time-property-valuations-how-ai-algorithms-are-making-it-possible/

5. Veres, O., Ilchuk, P., & Kots, O. (2021). *Data Science Methods in Project Financing Involvement*, In 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT) (Vol. 2, pp. 411–414). DOI: 10.1109/CSIT52700.2021.9648679

6. Veres, O., Ilchuk, P., & Kots, O. (2023). *Data Analytics on Debt Financing Research Based on Scopus and WoS Metrics*, In 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT). DOI: 10.1109/CSIT61576.2023.10324179

7. (n. d.). *DIM.RIA – all real estate of Ukraine. Sale and rent of any real estate*. Dom.ria.com. https://dom.ria.com/uk/

8. (n. d.). *Ukrainian classifieds service*. Olx.ua. https://www.olx.ua/uk/nedvizhimost/

9. (n. d.). *Agents. Tours. Loans. Homes.* Zillow. https://www.zillow.com/

10. (n. d.). *for Sale, Real Estate & Property Listing*. Realtor.com. https://www.realtor.com/

11. (n. d.). *Real Estate, Homes for Sale, MLS Listings, Agents*. Redfin.com. https://www.redfin.com/

12. (n. d.). *Most Trusted Provider of Real Estate Information*. Propstream.com. https://www.propstream.com/

13. Berezhna, N. (2021). Buying a home: do you need a realtor and how much do his services cost in Ukraine. URL: https://realestate.24tv.ua/kupivlya-zhitla-potriben-rieltor-skilki-koshtuyut-ostanni-novini_n1525065

14. (n. d.). *How much is my home worth?* Zillow.com. https://www.zillow.com/how-much-is-my-home-worth/

15. (n. d.) *Machine Learning Regression Explained*. Seldon.io. https://www.seldon.io/machine-learning-regression-explained

16. Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

17. Finch, W. H., Bolin, J. E., & Kelley, K. (2019). *Multilevel modeling using R*. Crc Press

18. Evans, C., Leckie, G., & Merlo, J. (2020). Multilevel versus SingleLevel Regression for the Analysis of Multilevel Information: The Case of Quantitative Intersectional Analysis. *Social Science and Medicine*, 245, 112499. Article 112499. https://doi.org/10.1016/j.socscimed.2019.112499

19. (n. d.). *What is Simple Linear Regression in Machine Learning?* Simplilearn.com. https://www.simplilearn.com/what-is-simple-linear-regression-in-machine-learning-article

20. Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140-147.

21. Polzer, D. (2021, June 21). *7 of the Most Used Regression Algorithms and How to Choose the Right One. Linear and Polynomial Regression, RANSAC, Decision Tree, Random Forest, Gaussian Process and Support Vector Regression*. Towardsdatascience.com. https://towardsdatascience.com/7-of-the-most-commonly-used-regression-algorithms-and-how-to-choose-the-right-one-fc3c8890f9e3

22. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

23. Dawson, C. (2021, January 23). *Understanding Multiple Linear Regression*. Medium.com. https://medium.com/swlh/understanding-multiple-linear-regression-e0a93327e960

24. Mahaboob, B., Praveen, J. P., Rao, B. A., Haranadh, Y., Narayana, C., & Prakash, G. B. (2020). A study on multiple linear regression using matrix calculus. *Advancecs in Mathematics Scientifc journal*, *9*(7), 1–10.

25. Bouzebda, S., Souddi, Y., & Madani, F. (2024). Weak Convergence of the Conditional Set-Indexed Empirical Process for Missing at Random Functional Ergodic Data. *Mathematics*, *12*(3), 448.

26. Zhou, Y., & He, D. (2024). Multi-Target Feature Selection with Adaptive Graph Learning and Target Correlations. *Mathematics*, *12*(3), 372.

27. Li, T., Frank, K. A., & Chen, M. (2024). A Conceptual Framework for Quantifying the Robustness of a Regression-Based Causal Inference in Observational Study. *Mathematics*, *12*(3), 388.

28. Leyland, A. H., & Groenewegen, P. P. (2020). *Multilevel Modelling for Public Health and Health Services Research: Health in Context* (p. 286). Springer Nature. https://doi.org/10.1007/978-3-030-34801-4.

29. (n. d.) *Measure of impurity*. Medium.com. https://medium.com/@viswatejaster/measure-of-impurity-62bda86d8760

30. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967.

31. Rishit, P. (2023). *Understanding K-Nearest Neighbors: A Simple Approach to Classification and Regression. Demystifying K-Nearest Neighbors: Unveiling the Power of Proximity-based Algorithm*. Pub.Towardsai.net. https://pub.towardsai.net/understanding-k-nearest-neighbors-a-simple-approach-to-classification-and-regression-e4b30b37f151

32. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

33. Bhandari, P. (2020, June 30). *Central Tendency | Understanding the Mean, Median & Mode*. Scribbr.com. https://www.scribbr.com/statistics/central-tendency/

34. (n. d.). *What Is R Squared And Negative R Squared*. Fairlynerdy.com. http://www.fairlynerdy.com/what-is-r-squared/

35. (n. d.). *ML | Introduction to Data in Machine Learning*. Geeksforgeeks.org. https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/

36. (n. d.). *What are the average prices on the secondary housing market in Ukraine: how much will you have to pay for a one-room apartment.* Sud.ua. https:https://sud.ua/uk/news/ukraine/259841-kakie-srednie-tseny-na-vtorichnom-rynke-zhilya-po-ukraine-skolko-pridetsya-otdat-za-odnokomnatnuyu-kvartiru

37. (n. d.). *Real estate Kyiv and region*. https://t.me/ppbestate

38. (n. d.). *Real estate of the Kyiv region*. https://t.me/Neruhomist_Kyiv_region

39. (n. d.). *Telethon's Documentation*. Docs.Telethon.dev. https://docs.telethon.dev/en/stable/

40. (n. d.). *Nominatim 4.3.0 Manual*. Nominatim.org. https://nominatim.org/release-docs/latest/api/Overview/

41. (n. d.). *Welcome to Python Overpass API's doc*

42. *umentation!* Python-Overpy.Readthedocs.io. https://python-overpy.readthedocs.io/en/latest/

FORECASTING THE VALUE OF REAL ESTATE USING MACHINE LEARNING TOOLS

Oleh Veres[1], Andrii Shymoniak [2]

Lviv Polytechnic National University, Information Systems and Networks Department, Lviv, Ukraine
[1] Oleh.M.Veres@lpnu.ua, ORCID 0000-0001-9149-4752
[2] andrii.shymoniak.msaad.2022@lpnu.ua, ORCID 0009-0003-7068-5818

Correct valuation of real estate plays a crucial role in the process of buying and selling. We have carefully studied the existing applications with which we carry out real estate transactions, described their features, advantages and disadvantages. The developed model will help sellers get an estimate of their property according to the parameters entered, which can serve as a starting point for establishing

the final value. The computation of real estate values has historically been based primarily on the method of analyzing data manually and subjective estimates, often resulting in errors and delays. The use of machine learning algorithms in solving this problem turned out to be effective, since it has a number of advantages over the manual estimation method, namely: a high level of accuracy, elimination of subjectivity and bias in estimates, time efficiency, cost reduction, use of geospatial data and substantiation of results. The process of creating a machine learning model is conditionally decomposed into four stages, which include collecting data, filtering, processing, supplementing, dividing into different samples and training the model based on this data. We considered the most popular regression algorithms, briefly described the principle of their work, as well as metrics with which you can evaluate the quality of the predicted values of the models. Standard parameters were used to test linear regression algorithms, decision tree, nearest neighbor method, support vector method, and random forest. The determination coefficient R-square is chosen as the main metric. Comparing the coefficient of determination of the results, it became clear that the algorithm "random forest" showed the best result. Having manually selected hyper parameters for this algorithm, the average value of the absolute error of the predicted value is 8.49 %, and the median is 1.9 %. The constructed model meets the established quality requirements and is ready for implementation in the information system of forecasting the value of real estate. For buyers, this service will be relevant, since they will be able to search for real estate according to the parameters entered by them, which have a favorable price for the purchase.

Key words: analysis; machine learning; regression.