

УДК 004.94

DOI <https://doi.org/10.36994/2788-5518-2023-02-06-03>

АРХІТЕКТУРА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ ЗБОРУ Й АНАЛІЗУ ДАНИХ ПРО ВАРТІСТЬ ЗЕМЛІ СІЛЬСЬКОГОСПОДАРСЬКОГО ПРИЗНАЧЕННЯ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ

Бутенко М. Є., Відкритий міжнародний університет розвитку людини «Україна», Київ, Україна
<https://orcid.org/0009-0009-9545-6572>, maksym.butenko.study@gmail.com

Павленко В. І., канд. фіз.-мат. наук, доцент,
Відкритий міжнародний університет розвитку людини «Україна», Київ, Україна
<https://orcid.org/0000-0002-3958-0415>, pavlenko.v@i.ua

Анотація. У роботі розглянуто загальний підхід до архітектури програмного забезпечення, збору й аналізу даних за допомогою методів машинного навчання для передбачення вартості землі сільськогосподарського призначення. Наведено нові результати дослідження з використання методів LightGBM, Fast Tree, Fast Forest. Виділені етапи: збір даних, зберігання, дослідження та трансформація, агрегування, оптимізація та використання методів машинного навчання. Запропоновано використання декількох програмних застосунків для імплементації перерахованих етапів. Наведена архітектура взаємодії таких застосунків. Оглянуто використання даних про інфляцію та їх вплив на якість передбачення. До даних, зібраних на майданчиках з аукціонів, сайтів оголошень (перше джерело даних), додані дані із сайту Державної служби України з питань геодезії, картографії та кадастру (друге джерело даних). Набір даних, зібраний з аукціонів і оголошень, становить 2 123 рядки. З даних, наявних на сайті Державної служби України з питань геодезії, картографії та кадастру, відібрано інформацію про 102 924 ділянок. З першого джерела виділено повний набір усіх даних D1; набори даних D2 – множина D1 без даних про оренду та D3 – множина D1, де врахована інфляція за даними Національного банку України. З другого джерела виділено набір D4 – ділянки з типом «договір купівлі-продажу» за 2022–2023 роки; D5 – підмножина D4 для ділянок, вартість за 1 га яких не менше ніж 10 000 грн та не більше ніж 150 000 грн; D6 – множина D5, до якої додана інформація про відстані від великих міст України, та D7 – множина D6 з урахуванням інфляції. Використано методи машинного навчання для наборів даних, зібраних із цих джерел. Передбачення вартості з похибкою не більше за 10% вважається коректним. Результатом якості роботи машинного навчання залежно від алгоритму та набору даних є значення коефіцієнта детермінації до 0,897 та з відсотком коректних передбачень до 91. Для другого джерела точність оцінки гірша, ніж для першого, через меншу кількість атрибутів і відсутність інформації про оренду, термін оренди, родючість ґрунтів, деякі неточності у даних. Дослідження показують, що покращення якості початкового набору даних (збільшення повноти, репрезентативності) значно впливає на коефіцієнт детермінації та відсоток коректних передбачень навченої моделі.

Ключові слова: машинне навчання, передбачення вартості землі, прогнозне моделювання, методу LightGBM, Fast Tree, Fast Forest.

SOFTWARE ARCHITECTURE FOR THE COLLECTION AND ANALYSIS OF AGRICULTURAL LAND PRICE DATA USING MACHINE LEARNING

Maksym Butenko, Open International University of Human Development "Ukraine", Kyiv, Ukraine
<https://orcid.org/0009-0009-9545-6572>, maksym.butenko.study@gmail.com

Volodymyr Pavlenko, Ph.D., Ass. Prof. Open International University of Human Development "Ukraine", Kyiv, Ukraine
<https://orcid.org/0000-0002-3958-0415>, pavlenko.v@i.ua

Abstract. The work presents a general approach to the software architecture for data collection and analysis using machine learning methods to predict the price of agricultural land. A brief review of recently published literature is made. New research results using LightGBM, Fast Tree, and Fast Forest methods to predict land price are provided. The highlighted stages include data collection, storage, exploration, and transformation, aggregation, optimization, and the use of machine learning methods. It is proposed to use several software applications for the implementation of these stages. An architecture diagram of the interaction of such applications is provided. The use of inflation data and the impact of such data on forecast quality reviewed. In addition to the dataset collected from the auctions and classified sites (first data source), data available on the website of the State Service of Ukraine for Geodesy, Cartography,

and Cadastre (second data source) were used. The dataset collected from first data source consist of 2123 rows, collected from seconds data source contains 815529 (further filtered to less rows due to different types of contract, year, etc). List of datasets for which machine learning was performed has been modified: attribute about soil type, distances from big cities, inflation were added. Information from the data available on the website of the State Service of Ukraine for Geodesy, Cartography, and Cadastre was converted to dataset, and it was further filtered and utilized, ranging from 75843 to 102924 rows. From the initial data set (D1), two additional ones were created: without data on land rent (D2) and with prices that take into account inflation (according to the National Bank of Ukraine) (D3). Experiments on removing the value of the regulatory monetary assessment of land reviewed. Also, for data source data was converted to four datasets. Dataset D4 – constructed using date filters for 2022 and 2023 for the "sale-purchase agreement" type. Rows without specified prices will be removed from the dataset. The total number of rows in dataset D4 is 102924. Dataset D5 – additionally apply a filter on the price per hectare of land, ranging from no less than 10000 UAH to no more than 150000 UAH. After this filtration, only 82241 parcels will remain in the dataset. Dataset D6 – expand dataset D5 with information about distances from the largest cities in Ukraine to the land. Dataset D7 – further expand dataset D6 with information about inflation. Thus, machine learning methods LightGBM, Fast Tree, and Fast Forest have been used for datasets collected from two different sources. If the cost prediction has an error of no more than 10%, it is considered correct. The result of the machine learning, depending on the algorithm and data set, for R^2 up to 0.897 and percentage of correct predictions up to 91%. Research shows that as the number of rows in the initial dataset increases, the quality of prediction improves. As for second data source – it have slightly worsen prediction results then first data source, but second data source usage have shown how fast dataset can be extended with new parameters (e.g. soil type, distances from cities, inflation) to significantly increase prediction quality. Another possible way to improve the quality of the forecast could be the integration of certain data sets on non-market factors affecting land value (e.g., a map of potential contamination by explosive objects). The summary suggests further steps to improve the quality of predictions.

Key words: machine learning, land price prediction, prediction modeling, LightGBM, Fast Tree, Fast Forest methods.

Вступ

У статті [1] показано важливість використання методів машинного навчання для передбачення вартості земельних ділянок сільськогосподарського призначення в Україні. Особливо це актуально у зв'язку з важливими змінами в законах України щодо умов купівлі-продажу земельних ділянок. У цій статті пропонується загальний підхід до архітектури програмного забезпечення зі збору й аналізу даних методами машинного навчання з урахуванням впливу інфляції на передбачення вартості. Додатково використані дані Державної служби України з питань геодезії, картографії та кадастру «Інформація про здійснені транзакції по земельним ділянкам за період» [2].

Коректне передбачення визначається за умовою [3]:

$$K = \left| \frac{x - x_i}{x_i} \right| \leq 0,1, \quad (1)$$

де K – передбачення, x – передбачене значення, x_i – реальне значення.

Відсоток коректних передбачень визначається за формулою

$$P = \frac{K_k}{K_a} * 100\%, \quad (2)$$

де P – відсоток коректних передбачень, K_k – кількість коректних передбачень, K_a – кількість передбачень усього.

Результатом машинного навчання в роботі [1] отримано моделі оцінки вартості земельних ділянок, що мають коефіцієнт детермінації 0,818 і дають 77% коректних передбачень для тестового набору даних (92% для тренувального набору даних).

У роботі [4] розглянуто питання оцінки вартості землі в Республіці Шрі-Ланка за допомогою методів машинного навчання: Multiple Linear Regression, Random Forest, Support Vector Regression, Extra Tree Regression, XGBoost. На відміну від робіт [3, 5] (ці роботи були детально розглянуті в [1]), у цій роботі не вираховується певний відсоток коректних передбачень, але наводяться результати роботи алгоритмів у метриках MSE (середньоквадратична похибка), RMSE (коренева середньоквадратична похибка), MAPE (середня абсолютна похибка передбачення), R^2 (коефіцієнт детермінації), де R^2 досягає значень від 0,87 до 0,9 при MSE від 0.09 до 0.12 та RMSE від 0.31 до 0.35. Найкращі результати показало використання методу Random Forest та XGBoost.

У роботі [6] обговорюється передбачення вартості побудови «зелених будинків» за допомогою методів машинного навчання, і описані підходи можуть бути корисні для оцінки вартості земельних ділянок. Стаття детально описує загальний підхід до вирішення такої задачі, а саме які дані та яким

чином треба збирати, наведено методологію передбачення вартості за допомогою блок-схеми. Методологія подібна до запропонованої у [7] та в цій статті. У роботі порівнюється використання методів машинного навчання XGBoost, Random Forest, Deep Neural Network (DNN). Найкращим результатом роботи є модель передбачення з коефіцієнтом детермінації R^2 на рівні 0,96. Результат використання методу машинного навчання XGBoost був кращим за результат використання DNN ($R^2 = 0,96$ для XGBoost та $R^2 = 0,95$ для DNN). Автори доходять висновку, що використання методів машинного навчання дає змогу вирішувати поставлену задачу.

У роботі [8] описано підхід до використання методів Gradient boosting, G-RNN, Random Forest, SVR, ARIMA до передбачення вартості продукції сільськогосподарської діяльності. Як і для ринку продажу земельних ділянок, існують певні проблеми збору даних. Автори описують загальний підхід до збору даних та їх приведення до вигляду, необхідного для машинного навчання. Через наявність історичних даних за певний проміжок часу використано алгоритми аналізу часових рядів. У висновках зазначається, що проблема такого аналізу є складною та потребує якісних початкових даних, але й отримані результати можуть бути додатковим чинником для вирішення питань про плани на продаж продукції для господарств.

Джерела даних та архітектура програмного забезпечення

Розробка архітектури програмного забезпечення для навчання за допомогою методів машинного навчання є важливою для забезпечення ефективності, надійності та масштабованості аналізу даних. Використання простих алгоритмів без адекватної архітектури може призвести до ряду проблем, включно з низькою точністю моделей, складністю збору даних та їх зберігання, повільною обробкою великих обсягів даних, важкістю підтримки й розширення функціонала програмного забезпечення з плином часу. Розробка архітектури дає змогу структурувати процеси збору, обробки й аналізу даних, забезпечуючи оптимальне використання ресурсів і підвищуючи ефективність моделей машинного навчання. Правильно побудована архітектура сприяє створенню гнучких систем, які можна налаштовувати й розширювати відповідно до потреб користувачів. У статті [1] показано, що перевагами використання методів машинного навчання є прискорення та здешевлення оцінки вартості землі.

У роботі [7] описані такі етапи (потреби) для успішного аналізу даних за допомогою алгоритмів машинного навчання: збір даних (collect), переміщення та зберігання (move/store), дослідження та трансформація (explore/transform), агрегування та розбивка (aggregate/label), вивчення й оптимізація даних (learn/optimize), машинне навчання.

Збір, переміщення та зберігання даних, агрегування та розбивка

Збір: як описано в [1], існує декілька можливих джерел даних – державні реєстри, інформація, яка доступна на сайтах оголошень чи аукціонів тощо. На цьому етапі визначимо джерела, інструментарій для збору даних, необхідність отримання звітів і журналів щодо автоматичного збору даних. Виконаємо автоматично збір даних окремо з кожного джерела, якщо вони мають відмінності у форматі збереження даних, частоти оновлення, об'єму тощо.

Переміщення та зберігання: зібрані дані збережемо в локальному середовищі. Для даних у файловому форматі csv (файл із даними, розділеними комою чи символом) є два можливі сценарії зберігання. Перший – завантажити в оперативну пам'ять, де виконуватиметься машинне навчання; другий – перенести інформацію з файлу в локальне сховище. Перевагою переносу даних в певну базу є наявність потужних можливостей щодо роботи з даними в сучасних СКБД. Для зберігання даних зручно використовувати бази даних, що підтримують SQL, або бази NoSQL. У випадку NoSQL код для інтеграції даних здається дещо простішим (через відсутність обов'язкової типізації, створення таблиць), але потребує контролю за джерелом даних, тому що набір даних (атрибутів, що описують земельну ділянку) змінюється, а отже, модифікується код, що збирає дані.

Дослідження та трансформація: проаналізуємо зібрані дані, їх структуру, потенційні взаємозв'язки, перевіримо на достовірність, видалимо некоректні дані (приміром, ціна продажу менша за нормативно-грошову оцінку землі, а на цей час такі продажі недозволені). Далі трансформуємо дані у зручний для подальшого використання формат: приведемо до єдиного формату вартість, дати покупки-продажу ділянки. Розширимо отриманий набір даними, які суттєво впливають на вартість землі (типом ґрунту, від якого залежить родючість землі, відстанню від великих міст), врахуємо інфляцію на дату оцінки. Встановлення типу ґрунту автоматизується простим алгоритмом, який визначає колір за певними координатами на карті й перетворює його у відповідний атрибут «тип ґрунту». Зручно користуватися кольоровою картою «Ґрунти України» [9] для визначення типу ґрунтів через її високу контрастність. Обрахуємо в кілометрах відстані від великих міст (адміністративних центрів областей) до ділянки за допомогою програмного коду, який вираховує відстань за формулою [10]:

$$\text{distance} = \arccos(\sin(\text{latitude}_1) * \sin(\text{latitude}_2) + \cos(\text{latitude}_1) * \cos(\text{latitude}_2) * \cos(\text{longitude}_1 - \text{longitude}_2)) * 6371 \quad (3)$$

де 6371 – це радіус Землі в км, distance – значення відстані, latitude та longitude – відповідні значення координат.

Обрахунок нової ціни ділянки з урахуванням інфляції від моменту продажу розрахуємо за формулою:

$$Price_{new} = Price_{old} * \left(1 + \sum_{k=2022}^{2023} days_k * avgInfPerDay_k \right) \quad (4)$$

де $Price_{new}$ – нова ціна для машинного навчання, $Price_{old}$ – ціна на дату реєстрації права на земельну ділянку, $days_k$ – кількість днів, які вже пройшли в k -му році, $avgInfPerDay_k$ – середня інфляція за один день у k -му році. Значення інфляції візьмемо з [11].

Виконаємо категорійне кодування: переведемо текстові значення атрибутів у кількісні в числовому форматі. Із двох широко відомих технік категорійного кодування: кодування міток (label encoding) та кодування в унітарному коді (one hot encoding) – виберемо останнє, оскільки складніше зіставляти атрибути, задані у строковій формі.

Агрегування та розбивка: вибрані дані збираємо в єдиний набір і розбиваємо на тренувальний і тестовий.

Вивчення й оптимізація даних: досліджуємо використання різних методів машинного навчання, встановлюємо впливові атрибути для моделі.

За наявності N джерел даних на рис. 1 зображено відносини між застосунками, потрібними для реалізації вказаних вище етапів. Джерела даних $\{S_1, S_2, S_3, \dots, S_N\}$ вимагають в загальному випадку N застосунків. Під час збирання даних ці застосунки виконують перевірку коректності ціни, площі ділянки в припустимих межах тощо. Також на рисунку зображено різні локальні сховища, де зберігаються дані з різних джерел (кількість сховищ та їх зв'язок із застосунками для збору даних залежить від особливостей програмної реалізації). Застосунок для агрегування використовує дані з локальних сховищ і створює єдиний набір із додатковими атрибутами (тип ґрунту, відстань від великих міст та інфляція за час до моменту оцінки). Застосунок для навчання використовує цей набір для отримання моделі обрахунку вартості земельної ділянки у вигляді, доступному для подальшого використання.

Якщо через користувацький інтерфейс програмного забезпечення ввести атрибути будь-якої ділянки, то програма покаже її вартість, яка є оціночною вартістю.

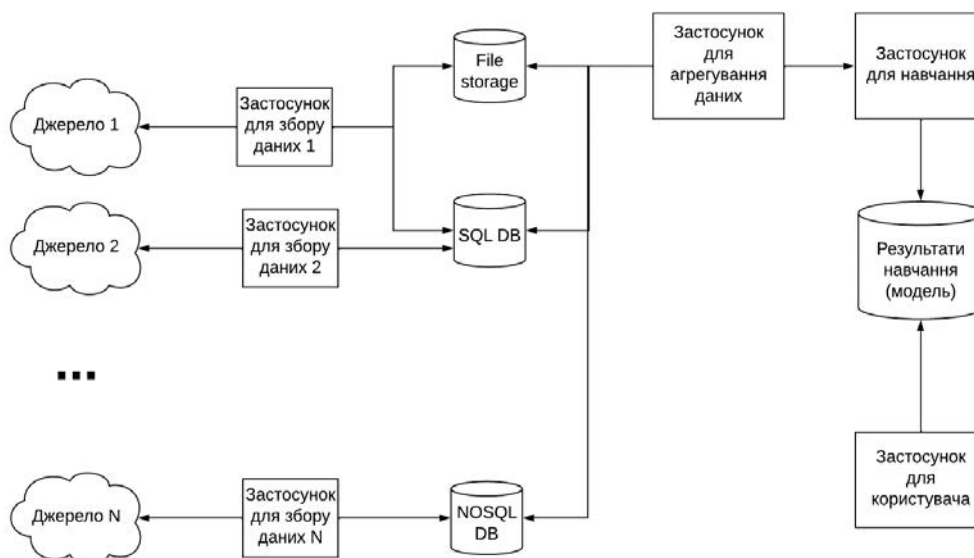


Рис. 1. Загальна архітектура програмного забезпечення для збору й аналізу даних про вартість землі сільськогосподарського призначення з використанням машинного навчання

Наведена архітектура дає змогу використовувати повністю автоматизовані застосунки для збору даних. У разі потреби можна додавати дані вручну за допомогою користувацького інтерфейсу.

Схожим чином, як у статті [1], зібрані дані з аукціонів, сайтів оголошень, тощо (перше джерело даних). Для 2 123 ділянок зібрано інформацію, що містить їх вартість, вартість оренди, термін оренди, НГО (нормативно-грошову оцінку), кадастровий номер, адміністративну приналежність, географічні координати, площу, прогнозовану родючість, тип ґрунту. Інші дані, які описують інформацію щодо продажу земельних ділянок, також є на сайті [2] (друге джерело даних). Станом на 01.01.2024 файл містить дані про 815 529 транзакцій, а саме: кадастровий номер земельної ділянки, КОАТУУ (класифікатор об'єктів адміністративно-територіального устрою України), площу ділянки, область, район, населений пункт, вулицю, назву територіальної громади, цільове призначення, назву угоддя, тип угоди. Цей перелік полів менший, ніж доступно в першому джерелі даних: відсутня інформація

про оренду, тип ґрунту та родючість. У файлі частина даних належить до транзакцій щодо спадщини й інших випадків, коли частина транзакцій не є купівлею-продажем ділянки. Для 197 886 ділянок із типом угоди купівлі-продажу лише у 137 358 випадках вказано ціну продажу. Ціна деяких ділянок або висока, або занадто низька. Тобто, можливо, існують певні особливості ділянки (нааявність корисних копалин із дозволом держави на їх видобуток тощо). За 2022 рік у файлі доступна інформація про 21 287 угод купівлі-продажу, а за 2023 рік – про 81 646 таких угод.

Створимо такі набори даних із двох джерел:

– з першого джерела виділимо повний набір усіх даних D1; D2 – множина D1 без даних про оренду, D3 – множина D1 з урахуванням інфляції в ціні земельної ділянки. Загальна кількість рядків у кожному наборі даних D1–D3 – 2 123 (1 486 у тренувальному та 637 у тестовому);

– із другого джерела виділимо такі набори даних: D4 – підмножина повного набору із застосуванням фільтрів за датою за 2022 та 2023 роки, типу «договір купівлі-продажу», і тільки ділянки, для яких вказана ціна продажу. Загальна кількість рядків у наборі даних D4 – 102 924; D5 – підмножина D4 для ділянок, вартість за 1 га яких не менше ніж 10 000 грн та не більше ніж 150 000 грн. Кількість ділянок у наборі D5 становить 82 241. D6 – множина D5, розширена інформацією про відстані від найбільших міст України до земельної ділянки. D7 – множина D6, розширена інформацією про інфляцію.

Кожен набір даних розбитий на два: тренувальний і тестовий у співвідношенні 7 : 3. Дані відсортовані випадковим чином, але в різних наборах даних збережене розподілення між тренувальними та тестовими даними. Тобто, якщо у наборі D1 певна ділянка потрапила до тренувального набору даних, то і в D2 ця ділянка буде в тренувальному наборі.

Виконання досліджень

З використанням 7 наборів даних, які виділені з двох джерел, проведемо машинне навчання. Відмінністю в порівнянні з [1] буде врахування даних про інфляцію та використання даних від Державної служби України з питань геодезії, картографії та кадастру [2].

Як мову програмування виберемо мову C# та фреймворк машинного навчання ML.NET. У цьому фреймворку доступні методи машинного навчання LightGBM (цей метод, як і XGBoost, належить до методів градієнтного бустінгу – підсилення), Fast Tree, Fast Forest.

Проведемо навчання, оптимізуючи моделі за коефіцієнтом детермінації (R^2).

Результати навчання наведено в таблиці 1.

Таблиця 1
Результат машинного навчання, коефіцієнт детермінації

Номер набору даних	R^2		
	LightGBM	Fast Tree	Fast Forest
D1	0,851	0,897	0,648
D2	0,543	0,58	0,56
D3	0,874	0,862	0,660
D4	0,003	0,006	0,004
D5	0,259	0,345	0,294
D6	0,314	0,328	0,295
D7	0,337	0,368	0,301

Для визначення впливовості певного атрибуту в наборі даних розрахуємо значення важливості індексу перестановки PFI (Permutation Feature Importance). У таблиці 2 наведено впливовість атрибутів для набору даних D1. Значення впливовості дорівнює індексу перестановки.

Таблиця 2
Впливовість атрибутів для набору даних D1

Індекс	Впливовість атрибута		
	LightGBM	Fast Tree	Fast Forest
1	Area (0,84)	area (1,06)	dystanceFromOdessa (0,28)
2	rentalYield (0,8)	rentalYield (0,96)	rentRate (0,13)
3	rentRate (0,77)	rentRate (0,92)	rentalYield (0,11)
4	dystanceFromOdessa (0,26)	dystanceFromOdessa (0,11)	coordinateLatitude (0,09)
5	dystanceFromKrivoiRog (0,09)	distanceFromKyiv (0,04)	distanceFromKyiv (0,04)

На рис. 2 відображено відношення PFI для п'яти найвпливовіших атрибутів для LightGBM.

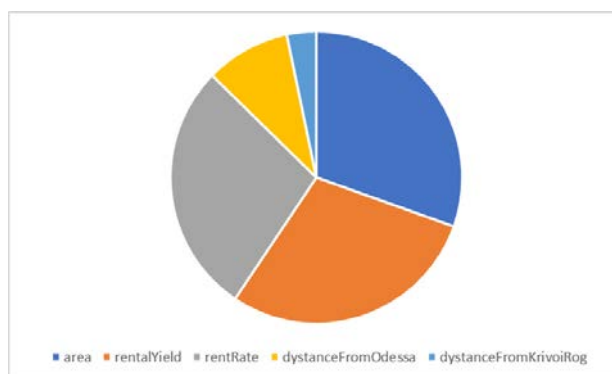


Рис. 2. Найвпливовіші атрибути для LightGBM

На рис. 3 відображено відношення PFI для п'яти найвпливовіших атрибутів для Fast Tree.

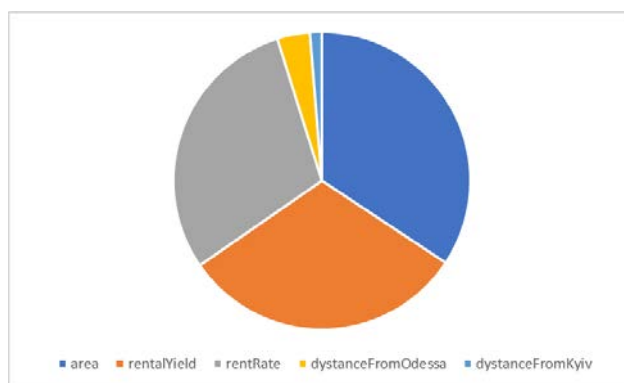


Рис. 3. Найвпливовіші атрибути для Fast Tree

На рис. 4 відображено відношення PFI для п'яти найвпливовіших атрибутів для Fast Forest.

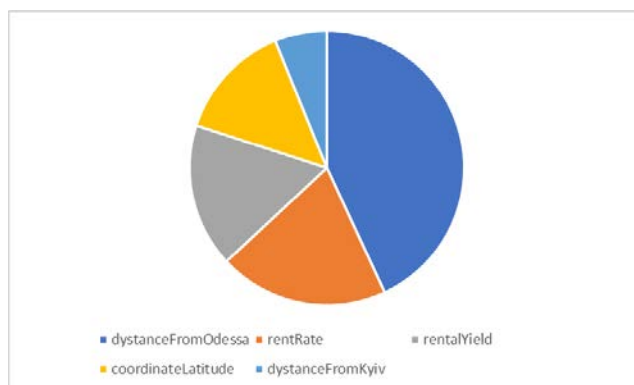


Рис. 4. Найвпливовіші атрибути для Fast Forest

У таблицях 3–5 наведено відсоток коректних передбачень згідно з формулою (2) для тестового набору даних.

Таблиця 3
Відсоток коректних передбачень – LightGBM

Номер набору даних	D1	D2	D3	D4	D5	D6	D7
Відсоток коректний передбачень	84	55	88	2	21	18	59

Таблиця 4
Відсоток коректних передбачень – Fast Tree

Номер набору даних	D1	D2	D3	D4	D5	D6	D7
Відсоток коректний передбачень	89	38	91	23	29	15	59

Таблиця 5
Відсоток коректних передбачень – Fast Forest

Номер набору даних	D1	D2	D3	D4	D5	D6	D7
Відсоток коректний передбачень	43	36	47	9	13	8	56

Наведені результати показують, що на якість моделі значно впливає наявність інформації про оренду (D2 має гірші результати, ніж D1). Врахування інфляції за два роки покращує значення $\llbracket \text{Eqn022.eps} \rrbracket$ на 2–4% для першого набору даних і на 2–11% – для другого. З додаванням даних про відстані від великих міст і з урахуванням рівня інфляції для другого набору даних якість моделі покращується в декілька разів (D7 має набагато кращий відсоток коректних передбачень, ніж D4). При цьому додавання лише інформації про відстані чи лише врахування інфляції не дає значного приросту коректного передбачення (порівняння D4 та D5, D4 та D6). У певних випадках (D1 і D3, метод Fast Tree; D5 і D6, метод LightGBM) коефіцієнт детермінації погіршується, а відсоток коректних передбачень незначно покращується. Якість моделі для другого джерела даних нижча, ніж для першого. Найгірші результати отримано для D4 через наявність даних, що здаються недостовірними і які були видалені для наборів D5–D7.

У наборах D3 та D7 621 ділянка має однаковий кадастровий номер. З них для моделі D3 і моделі D7 передбачено коректно 573 та 421 випадків відповідно, тобто відсоток коректних передбачень становив 92% та 67%. Цей відсоток кращий, ніж наведено в таблицях 3–5, бо частина із 621 ділянки є в навчальному наборі даних.

Висновки

У роботі наведено загальну схему архітектури програмного забезпечення для збору й аналізу даних про вартість землі сільськогосподарського призначення з використанням машинного навчання, яка охоплює чотири різні типи застосунків: для збору даних, для агрегування даних, для навчання та застосунок для користувача. Отримано результати роботи моделі передбачення для даних із двох джерел. Відсоток коректних передбачень залежно від набору даних і вибраного алгоритму сягає 91% для тестових наборів даних. Найкраще значення коефіцієнта детермінації залежно від набору даних і вибраного алгоритму становить 0,897, що є досить гарним результатом для подальшої роботи в цьому напрямі. Набір даних, що містить інформацію про ціни з урахуванням інфляції за два останні роки, на 4% краще передбачає вартість землі. Також, на відміну від [1], на перше місце за значенням $\llbracket \text{Eqn023.eps} \rrbracket$ вийшов алгоритм Fast Tree. Додавання інформації про відстані від великих міст до набору даних із другого джерела покращило коефіцієнт детермінації на 20% (з 0,259 до 0,314). Якість передбачення для другого джерела гірша через меншу кількість атрибутів і відсутність інформації про оренду, термін оренди, родючість ґрунтів, деякі неточності в даних (приміром, неринкова ціна за договором купівлі-продажу, тобто менша за НГО). Кращим результатом для другого джерела є коефіцієнт детермінації 0,368 із відсотком коректних передбачень 56%. Застосування другого джерела показує, що за допомогою додавання нових атрибутів (тип ґрунту, відстань до ділянки, інфляція) точність передбачення значно покращується.

У порівнянні з [1] отримано модель, яка краще відповідає поставленій задачі знаходження достовірної оцінки вартості земельної ділянки. Відсоток коректних передбачень для тестового набору даних збільшився із 77 до 91%, коефіцієнт детермінації зріс з 0,818 до 0,897.

Через те що на ціну можуть вплинути певні неринкові чинники (забрудненість ділянки вибухонебезпечними предметами і т. ін.), подальша інтеграція додаткових джерел даних (приміром, карти ДСНС України «Інтерактивна мапа територій, які потенційно можуть бути забруднені вибухонебезпечними предметами» [12]) підвищить точність оцінки.

Література

1. Бутенко М. Є., Павленко В. І. Передбачення вартості землі сільськогосподарського призначення за допомогою алгоритмів машинного навчання. Інфокомунікаційні та комп'ютерні технології. 2023. № 1. С. 161–167.
2. Інформація про здійснені транзакції по земельним ділянкам за період, розділ «Моніторинг земельних відносин». <https://land.gov.ua/>. URL: <https://land.gov.ua/> (дата звернення: 10.01.2024).
3. Machine-Learning-Based Prediction of Land Prices in Seoul, South Korea / K. Jungsun et al. Sustainability. 2021. No. 13 (23).
4. Naotunna R. A Model for the Estimation of Land Prices in Colombo District using Web Scraped Data. University of Colombo School of Computing. 2022.

5. Yunjong K., Seungwoo C., Mun Yong Y. Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices. *Sustainability*. 2020. No. 12 (14).
6. Extreme gradient boosting-based machine learning approach for green building cost prediction / O. Alshboul et al. *Sustainability*. 2022. No. 14.
7. Rogati M. The AI Hierarchy of Needs. <https://medium.com/hackernoon/the-ai-hierarchy-of-needs-18f111fcc007>. URL: <https://medium.com/hackernoon/the-ai-hierarchy-of-needs-18f111fcc007> (дата звернення: 11.03.2024).
8. Ranjit K. Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. *Plos one*. 2022. No. 17(7).
9. Ґрунти України. Карта. За ред. М. І. Полупан, В. Б. Соловей, В. А. Величко. Національний науковий центр «Інститут ґрунтознавства та агрохімії ім. О. Н. Соколовського», 2005.
10. Houde K.V., Hegadi R.S. Optimizing Drone Navigation Using Shortest Path Algorithms. *International Conference on Recent Trends in Image Processing and Pattern Recognition*. 2023.
11. Коментар Національного банку щодо рівня інфляції у 2022 році. <https://bank.gov.ua/ua/news/all/komentar-natsionalnogo-banku-schodo-rivnya-inflyatsiyi-u-2022-rotsi>. URL: <https://bank.gov.ua/ua/news/all/komentar-natsionalnogo-banku-schodo-rivnya-inflyatsiyi-u-2022-rotsi> (дата звернення: 11.03.2024).
12. Інтерактивна мапа територій, які потенційно можуть бути забруднені вибухонебезпечними предметами. <https://mine.dsns.gov.ua/>. URL: <https://mine.dsns.gov.ua/> (дата звернення: 11.03.2024).

References

- 1 Butenko, M., Pavlenko, V. (2023). Agricultural land price prediction using machine learning algorithms. *Infocommunication and computer technologies*. № 1. P. 161–167.
2. Information on transactions carried out on land for the period is available in the "Land Relations Monitoring" section on the website <https://land.gov.ua/>. [URL: <https://land.gov.ua/>] (date of access: 10.01.2024).
3. Machine-Learning-Based Prediction of Land Prices in Seoul, South Korea (2021). K. Jungsun et al. *Sustainability*. No. 13 (23).
4. Naotunna R. (2022). A Model for the Estimation of Land Prices in Colombo District using Web Scraped Data. *University of Colombo School of Computing*.
5. Yunjong, K., Seungwoo, C., Mun Yong, Y. (2020). Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices. *Sustainability*. No. 12 (14).
6. Extreme gradient boosting-based machine learning approach for green building cost prediction (2022). O. Alshboul et al. *Sustainability*. No. 14.
7. Rogati, M. The AI Hierarchy of Needs. <https://medium.com/hackernoon/the-ai-hierarchy-of-needs-18f111fcc007>. URL: <https://medium.com/hackernoon/the-ai-hierarchy-of-needs-18f111fcc007> (date of access: 11.03.2024).
8. Ranjit, K. (2022). Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. *Plos one*. No. 17 (7).
9. Soils of Ukraine. Map. (2005). Ed. M.I. Polupan, V.B. Solovey, V.A. Velychko. National Scientific Center "Institute of Soil Science and Agrochemistry named after O.N. Sokolovsky".
10. Houde, K.V., Hegadi, R.S. (2023). Optimizing Drone Navigation Using Shortest Path Algorithms. *International Conference on Recent Trends in Image Processing and Pattern Recognition*.
11. Commentary of the National Bank on the inflation level in 2022. URL: <https://bank.gov.ua/en/news/all/komentar-natsionalnogo-banku-schodo-rivnya-inflyatsiyi-u-2022-rotsi> (accessed: 11.03.2024).
12. Interactive map of areas potentially contaminated with explosive objects. <https://mine.dsns.gov.ua/>. URL: <https://mine.dsns.gov.ua/> (date of access: 11.03.2024).