

УДК 004.94

ПЕРЕДБАЧЕННЯ ВАРТОСТІ ЗЕМЛІ СІЛЬСЬКОГОСПОДАРСЬКОГО ПРИЗНАЧЕННЯ ЗА ДОПОМОГОЮ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

DOI 10.36994 / 2788-5518-2023-01-05-19

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

Павленко В. І., к.ф.-м.н., доц., Відкритий міжнародний університет розвитку людини «Україна», Інститут комп'ютерних технологій. Київ, Україна. <https://orcid.org/0000-0002-3958-0415>, pavlenko.v@i.ua

Анотація: Після запуску ринку землі в Україні питання визначення ринкової вартості землі стало більш нагальним. Використання методів машинного навчання дозволяє пришвидшити та спростити передбачення вартості. В даній роботі буде застосовано методи машинного навчання для передбачення вартості землі сільськогосподарського призначення в Україні. Існує декілька досліджень які описують цю задачу, але на прикладі інших країн. Набір даних для машинного навчання зібрано з мережі Інтернет і включає в себе 800 рядків (для деяких атрибутів є пропуски даних). Зроблено огляд проблем збору таких даних. В роботі передбачення вартості землі виконане за допомогою методів машинного навчання LightGBM, Fast Tree, Fast Forest. З початкового набору даних (D1) створено три додатковий: без даних про оренду землі (D2), без даних про нормативно-грошову оцінку (НГО) (D3) та без даних про оренду та НГО (D4). Використання методу LightGBM має найкращі результати передбачення, але у деяких випадках Fast Tree наближується до LightGBM по якості. Видалення НГО з набору даних для навчання покращує якість передбачення. Також, після додаткових розрахунків виявилось, що кореляція між ринковою вартістю та НГО відсутня. Також, на відміну від передбачення вартості нерухомості, використання методів градієнтного бустінгу надає кращі результати для передбачення вартості землі. Як і в інших статтях коректне передбачення визначено як передбачення, що має не більше 10% похибки. В залежності від набору даних та обраного метода машинного навчання якість передбачення складає від 35% до 92% коректних передбачень, що близько до результатів опублікованих в попередніх дослідженнях для інших країн. Оглянуто наявні обмеження дослідження та наведено можливі шляхи покращення якості передбачення вартості землі для наведених методів машинного навчання. Зокрема, подальше розширення набору даних (як по кількості рядків, так і по кількості атрибутів) повинно покращити якість передбачення.

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

AGRICULTURAL LAND PRICE PREDICTION USING MACHINE LEARNING ALGORITHMS.

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 University of Human Development "Ukraine", Kyiv, Ukraine. <https://orcid.org/0000-0002-3958-0415>, pavlenko.v@i.ua

Abstract: The paper considers the prediction of land value using LightGBM, Fast Tree, and Fast Forest machine learning methods. The training dataset is collected from the Internet and contains of 800 rows (in some cases there are data gaps for certain attributes). An overview of the problems of collecting such data is made. From the full dataset (D1), three additional ones were created: without data on land rent price (D2), without data on normative monetary assessment (NMA) (D3) and without data on rent price and NMA (D4). Using the LightGBM method has the best prediction results, but in some cases Fast Tree's prediction quality is similar to LightGBM. Removing NMA data from the dataset improves the quality of prediction, because after calculations it turned out that there is no correlation between market value and NMA. Also, unlike real estate value prediction, gradient boosting methods provides better results. A correct prediction is defined as a prediction with no more than 10% error, which follow similar approaches in other researches. Depending on the dataset and the chosen method of machine learning, the quality of prediction ranges from 35% to 92% of correct predictions. The paper describes research's limitations and possible ways to improve the quality of land price predictions for reviewed machine learning methods.

Keywords: *machine learning, land price prediction, prediction modeling, LightGBM, Fast Tree, Fast Forest methods.*

Вступ

Після змін у законі щодо обігу земель сільськогосподарського призначення питання визначення ринкової вартості землі в Україні стало більш нагальним [1]. Розділяють два основні підходи до визначення вартості землі – підхід, що ґрунтується на думці експерта-оцінювач та підхід, що вимагає застосування економіко-математичних та геоінформаційних моделей і алгоритмів. Другий підхід вважається більш точним, то ж було би цікаво використати машинне навчання для передбачення вартості землі сільськогосподарського призначення. В роботі Jungsun Kim, Jaewoong Won, Hyeongsoon Kim та Joonghyeok Heo [2] розглянуто питання використання алгоритмів машинного навчання для передбачення вартості землі в місті Сеул, і в цій статті буде розглянуто аналогічний підхід з використанням даних щодо вартості землі сільськогосподарського призначення в Україні.

Виконання досліджень

В статті [2] розглянута оцінка вартості землі на прикладі вартості землі не сільськогосподарського призначення у великому місті з використанням методів комп'ютерного навчання. Машинне навчання може використовувати багато атрибутів, та використовувати гнучкі специфікації моделей, що керуються даними [3]. Автори описали підхід до збору даних, що виокремив вартість землі і споруд на ній та розраховували передбачену вартість землі. В роботі використані алгоритми Random Forest (випадковий ліс) та XGBoost (реалізація підсилювального (boosting) методу машинного навчання). Автори пропонують визначити коректне передбачення вартості як передбачення що має не більше 10% похибки (і визначають що такий підхід співпадає з підходами запропонованими у [4,5]). Результатом роботи в залежності від фільтрів накладених на початковий набір даних та використаного алгоритму було від 54% до 87% коректно передбачених випадків.

В статі [3] розглянуто використання машинного навчання для реалізації методу порівняння продажів для передбачення вартості нерухомості. Як і в статті [2] автори використали підсилювальний (boosting) метод машинного навчання, але на відміну від XGBoost використали методи LightGBM та CatBoost. В якості метрик дослідження автори вибрали MAPE (середня абсолютна похибка передбачення, виражена у відсотках), RMSPE (середньоквадратична похибка передбачення, виражена у відсотках), і загальний відсоток коректних передбачень з таким самим критерієм як у статті [2] – похибка менше 10% вважається коректним передбаченням. Результатом роботи в залежності від набору даних та використаного алгоритму було від 62% до 89% коректно передбачених випадків.

В статті [6] розглянуті і підсумовані проблеми збору даних для використання алгоритмів машинного навчання для аналізу великих наборів даних в сільському господарстві (наприклад, передбачення урожаю, випадків появи захворювань, розпізнавання якості посаджених культур, тощо), в тому числі збору і аналізу даних про землю. В цілому доступ до таких даних є частково ускладненим через відсутність однієї загальнодоступної бази даних, і зазвичай збір даних про земельні ділянки вимагає використання даних з різних джерел.

Об'єктом дослідження є вартість землі. Площа України складає 603700 кв. км (трохи більше 60 мільйонів га), з яких 42.7 мільйони га – це землі сільськогосподарського призначення. Дані про ринкові пропозиції вартості ділянок землі доступні в Інтернеті, але зазвичай така інформація або не репрезентативна (має невеликий об'єм даних і не описує генеральну сукупність усіх ділянок). Наприклад сайти повідомлень, сайти аукціонів – мають інформацію лише про наявні ділянки), або має невеликий набір доступних атрибутів даних (наприклад, інформація з Державного земельного кадастру чи інформація про право власності та речові права на земельну ділянку має інформацію про власників і у деяких випадках про нормативно-грошову оцінку, але не містить інформації про якість ґрунту ділянки, тощо). Проблема збору даних описана в [6], то ж ця проблема частково існує і в інших країнах. Також, з 24.02.2023 і до завершення воєнного стану в Україні обмежено доступ до Національної структури геопросторових даних та Державного земельного кадастру, що також ускладнює збір даних для машинного навчання.

Використовуючи дані доступні у мережі Інтернет зібрано набір даних з 800 ділянок з наступними атрибутами (табл.1):

Таблиця 1.
Опис атрибутів набору даних

Назва змінної (атрибут)	Опис змінної	Опис значень
Залежні змінні		
Вартість 1 Га ділянки	Вартість у Грн	Середнє значення 56500
Незалежні змінні: географічні дані		
Місцезнаходження ділянки	Назва області, району	Наприклад, Сумська область
Географічні координати ділянки	Координати усіх кутових точок ділянки у вигляді широта та довгота.	
Площа ділянки	у Га	Середнє значення 2.9 Га
Прогнозована родючість ділянки	В залежності від типу ґрунту можна визначити приблизну родючість ділянки, у центнерах з Га	Наприклад, прогнозована родючість пшениці 52 ц/Га
Тип ґрунту	Виключено з набору даних для навчання, використовується для розрахунку значення прогнозованої родючості	Наприклад, "чорноземи типові". Дані наявні для 249 ділянок.
Незалежні змінні: дані про використання землі		
Вартість ділянки	Виключено з набору даних для навчання, використовується для розрахунку значення вартості за 1 Га	
Вартість оренди ділянки	Вартість у Грн за 1 рік	Середнє значення 10700 грн за рік за ділянку
Термін оренди	Термін у роках	Середнє значення 15.5 років
НГО (нормативно-грошова оцінка)	Грошова оцінка у Грн	571 ділянка не має інформації про НГО

Для деяких атрибутів табл.1 (наприклад прогнозована родючість, тип ґрунту) у зібраному наборі даних існують пропуски (відсутність даних).

В табл.2 наведено розподіл ділянок по регіонах.

Таблиця 2.
Розподіл по регіонах у наборі даних

Область	Кількість ділянок
Сумська	307
Чернігівська	130
Полтавська	98
Вінницька	82
Херсонська	80
Хмельницька	32
Черкаська	25
Київська	14
Харківська	11
Миколаївська	6
Івано-Франківська	4
Волинська	3
Луганська	2
Тернопільська	2
Рівненська	2
Донецька	1
Житомирська	1

В табл.3 надано приклад пропущених даних.

Таблиця 3.

Приклад наявності пропущених даних у наборі даних

Наявність даних про НГО	НГО вказаний
Так	229
Ні	571

Використовуючи зібрані дані, розширимо набір даних наступною інформацією: відстань у км від кожного обласного центра України, кількість точок у координатах ділянок (щоб визначити, чи є ділянка трикутною, чотирикутною чи багатокутною).

Для передбачення вартості землі дані були розбиті на чотири набори даних. Визначимо набір даних D1 як повний набір усіх зібраних даних, D2 як набір даних без даних про оренду, D3 як набір даних без даних про НГО, D4 як набір даних без даних про оренду та НГО. Також, випадковим чином розіб'ємо повний набір даних на 2 додаткових набори даних – тренувальний та тестовий у відношенні 7:3.

Для передбачення вартості землі використаємо мову програмування C# та фреймворк машинного навчання ML.NET. В передбаченні будуть використані методи навчання LightGBM (цей метод, як і XGBoost, відноситься до методів градієнтного бустінгу (підвищення)), Fast Tree, Fast Forest. Результати навчання будуть порівняні за допомогою таких метрик як: коефіцієнт детермінації, середньо-квадратична помилка та відсоток коректних передбачень (які будуть вираховані схожим чином як в [2,3] – якщо передбачення відрізняється від реальної ціни не більше ніж на 10% - воно вважається коректним).

Тобто, передбачення вважається коректним, якщо:

$$\left| \frac{\text{передбачене значення вартості} - \text{реальне значення вартості}}{\text{реальне значення вартості}} \right| \leq 0.1$$

Застосувавши навчання на протязі 120 секунд (навчання виконувалося на процесорі Intel Core i7-12700H) на тренувальному наборі даних отримуємо результати навчання, які наведені у табл.4 та табл.5.

Таблиця 4.

Результат машинного навчання, коефіцієнт детермінації

Номер набору даних	Коефіцієнт детермінації (RSquared)		
	LightGBM	Fast Tree	Fast Forest
D1	0.8182	0.72	0.5983
D2	0.4766	0.475	0.5046
D3	0.8092	0.8027	0.596
D4	0.4691	0.539	0.5051

Таблиця 5.

Результат машинного навчання, середньо-квадратична помилка

Номер набору даних	Середньо-квадратична помилка (RMS-loss)		
	LightGBM	Fast Tree	Fast Forest
D1	9663	11177	14124
D2	16192	15787	15764
D3	9831	10535	14155
D4	16587	15673	15778

Детальний опис метрик (RSquared, RMS-loss) моделі доступний у [7]. Як видно з результатів, найкращим можна вважати використання методу LightGBM.

Виконаємо розрахунок важливості атрибута за допомогою метода визначення важливості індексу перестановки (Permutation Feature Importance, PFI). Для можливості порівняння значень між різними методами машинного навчання приведемо значення важливості до 1 та визначимо 7 найважливіших атрибутів для кожного з методів. Важливість атрибутів для набору D1 в залежності від методу наведено у табл.6.

Таблиця 6.
Важливість атрибутів для набору даних D1

Пріоритет важливості	Важливість атрибута		
	LightGBM	Fast Tree	Fast Forest
1	Орендна плата (0.474)	Орендна плата (0.406)	Відсоток орендної плати (0.286)
2	Площа (0.251)	Площа (0.299)	Відстань від м. Вінниця (0.194)
3	Відсоток орендної плати (0.158)	Відсоток орендної плати (0.223)	Орендна плата (0.286)
4	Відстань від м. Вінниця (0.03)	Орендна плата після сплати податків (0.016)	Координата широти (0.098)
5	Координата широти (0.02)	Відстань від м. Вінниця (0.014)	Відстань від м. Одеса (0.064)
6	Відстань від м. Миколаїв (0.013)	Відстань від м. Київ (0.008)	Відстань від м. Харків (0.048)
7	Відстань від м. Харків (0.011)	Відстань від м. Чернівці (0.005)	Відстань від м. Житомир (0.043)

Як показано в табл.6, атрибут орендної плати має сильний вплив на результат передбачення. Без даних про орендну плату (D2) отримаємо важливість атрибутів що наведені у табл.7.

Таблиця 7
Важливість атрибутів для набору даних D2

Пріоритет важливості	Важливість атрибута		
	LightGBM	Fast Tree	Fast Forest
1	Відстань від м. Вінниця (0.219)	Відстань від м. Вінниця (0.254)	Відстань від м. Вінниця (0.262)
2	Площа (0.109)	Координата широти (0.165)	Координата широти (0.180)
3	Координата широти (0.088)	Площа (0.117)	Відстань від м. Одеса (0.115)
4	Відстань від м. Одеса (0.071)	Відстань від м. Харків (0.066)	Площа (0.065)
5	Відстань від м. Чернівці (0.067)	Відстань від м. Миколаїв (0.063)	Відстань від м. Київ (0.062)
6	Відстань від м. Харків (0.061)	Відстань від м. Одеса (0.037)	Відстань від м. Житомир (0.051)
7	Відстань від м. Суми (0.06)	Відстань від м. Київ (0.032)	Відстань від м. Харків (0.05)

Як показано в табл.7, атрибут про НГО не є важливим, бо в початковому наборі даних ці дані присутні лише у 229 випадках. Можна також припустити, що дані про НГО часто розходяться з ринковою вартістю землі, то ж додавання НГО не покращує результат машинного навчання. Якщо перевірити це за допомогою функції кореляції по даним про ринкову вартість землі та НГО, то отримаємо що кореляція дорівнює -0.0622, тобто припущення підтверджується.

В табл.8 наведено відсоток коректних передбачень. В залежності від набору даних, вибраного метода машинного навчання отримаємо від 35 до 92 відсотків коректних передбачень, що в цілому збігається з результатами отриманими в [2,3]. На наборі даних D1 найкращим методом був LightGBM, у той час як найгіршим – Fast Forest. При видаленні даних про НГО, Fast Tree метод значно покращує результат передбачення, то ж цей метод на наявному наборі даних має більші вимоги до якості даних та їх підготовки до машинного навчання.

Таблиця 8.
Відсоток коректних передбачень

	Відсоток коректних передбачень											
	LightGBM				Fast Tree				Fast Forest			
	1	2	3	4	1	2	3	4	1	2	3	4
Тренувальний набір даних	1.2	5.7	1.7	2.6	4.1	3.2	2.5	3.5	6.9	0	7.6	0
Тестовий набір даних	7.5	5.4	3.2	5	3.7	7	7	2	7.5	3	6.6	2.9
Повний набір даних	7.1	5.6	8.2	3.3	1	1.3	7.8	3.1	7.1	1.2	7.3	0.8

Щоб додатково порівняти результат передбачення за допомогою методів машинного навчання спробуємо визначити певну базову модель, модель без використання методів машинного навчання. Наприклад, таку модель можна визначити як – передбачення вартості робиться як середнє значення вартості землі ділянок що знаходяться поруч (наприклад, у радіусі 40 км). Такий підхід називається порівняльним аналізом ринку (Comparative Market Analysis) та оглянутий, наприклад у [10].

Результат використання такого передбачення наведено у табл.9, то ж ця модель гірша за будь який результат роботи машинного навчання.

Таблиця 9.
Результат передбачення для базової моделі

Середньо-квадратична помилка (RMS-loss), у Грн	17556.87
Кількість коректних передбачень	219
Кількість не-коректних передбачень	581
Відсоток коректний передбачень	27.35

Під час огляду цієї моделі також були виконані розрахунки для менших значень радіуса – у такому випадку середньо-квадратична помилка підвищувалася на 40-50% (до ~25000), а при збільшенні радіуса – результат передбачення майже не змінювався. Також, у деяких випадках ділянок поруч в наборі даних може не бути, тобто використання базового методу має ряд обмежень.

У статтях [8, 9] відзначено, що для передбачення вартості нерухомості кращі результати повинні давати методи Fast Forest. Результати даного дослідження показують, що використання

методу градієнтного бустінгу надає кращі результати для передбачення вартості землі ніж методи Fast Tree, Fast Forest. Це твердження вимагає додаткового дослідження, але в дослідженні [2] також зазначалася перевага методів градієнтного бустінгу над Fast Forest для передбачення вартості землі. Також, варто відмітити важливість підготовки даних та чутливість Fast Tree до наявності даних з пропусками. На прикладі використання набору даних D3 – якість передбачення зросла після того як було видалено поле з інформацією про НГО.

Обмеженнями даного дослідження можна відзначити дві основні проблеми – обмеження накладені початковим набором даних (D1) та складності побудови моделі машинного навчання, що буде надавати стабільне і якісне передбачення ціни довільної ділянки.

Обмеження початкового набору даних, такі як наявність лише 800 ділянок, деякі пропуски для певних ділянок можна покращувати. Також, як зазначалося вище – до кінця воєнного стану в Україні доступ до частини даних в державних реєстрах частково обмежений, то ж після закінчення воєнного стану кількість даних доступних для аналізу зросте. Можливим розширенням набору даних може бути інформація що не стосується безпосередньо земельної ділянки (наприклад, кількість жителів в найближчому населеному пункті, чи середня заробітна плата у регіоні, тощо). Складності використання машинного навчання, такі як отримання різних результатів при використанні різних методів машинного навчання, отримання різних значень важливості атрибутів, необхідність навчання з використанням різних гіперпараметрів методу, тощо також була відзначена в [2].

Висновки

В роботі було розглянуто використання методів машинного навчання для передбачення вартості землі. Використано три методи машинного навчання – LightGBM, Fast Tree, Fast Forest на наборі даних, зібраного з мережі Інтернет, що включає інформацію про 800 ділянок. Використання LightGBM має кращі результати ніж використання Fast Tree, Fast Forest, але у випадку видалення атрибутів що має пропуски (наприклад НГО) результат використання Fast Tree наближується до LightGBM. Результати дослідження можна продовжувати – покращення якості початкових даних та розширення кількості атрибутів повинно покращити точність передбачення.

Література

1. Черевко Г., Грисьо. М. ОЦІНКА І ЦІНА ЗЕМЛІ В УКРАЇНІ ТА ДОСВІД КРАЇН ЄВРОПИ / Г. Черевко, М. Грисьо. // Наукові записки Національного університету «Острозька академія». Серія «Економіка». – 2023. – №28. – С. 4–13.
2. Machine-Learning-Based Prediction of Land Prices in Seoul, South Korea / K.Jungsun, W. Jaewoong, K. Hyeongsoon, H. Joonghyeok. // Sustainability. – 2021. – №13(23).
3. Yunjong K. Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices / K. Yunjong, C. Seungwoo, Y. Mun Yong. // Sustainability. – 2020. – №12(14).
4. Crosby N. Property valuation variation and the 'margin of error' in the UK / N. Crosby, A. Lavers, J. Murdoch. // Journal of Property Research. – 1998. – №15. – С. 305–330.
5. Watkins C. The definition and identification of housing submarkets / Craig Watkins. // Environment and Planning. – 2001. – №33. – С. 2235 – 2253.
6. Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review / A.Cravero, S. Pardo, S. Sepulveda, L. Munoz. // 2022. – №12.
7. Evaluate your ML.NET model with metrics [Електронний ресурс] // microsoft.com. – 2023. – Режим доступу до ресурсу: <https://learn.microsoft.com/en-us/dotnet/machine-learning/resources/metrics>.
8. A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems / [J. Alfaro-Navarro, E. Cano, E. Alfaro-Cortes та ін.]. // Complexity. – 2020. – №2020. – С. 1–12.
9. Quang T. Housing Price Prediction via Improved Machine Learning Techniques / T. Quang, N. Minh, D. Hy. // Procedia Computer Science. – 2020. – №174. – С. 433–442.
10. What Is Comparative Market Analysis (CMA) in Real Estate? [Електронний ресурс] // Investopedia. – 2022. – Режим доступу до ресурсу: <https://www.investopedia.com/terms/c/comparative-market-analysis.asp>.

References

1. Cherevko H., Hryso. M. OTSINKA I TSINA ZEMLI V UKRAINI TA DOSVID KRAIN YEVIROPY / H. Cherevko, M. Hryso. // Naukovi zapysky Natsionalnoho universytetu «Ostrozka akademiia». Serii «Ekonomika». – 2023. – №28. – С. 4–13.2. Machine-Learning-Based Prediction of Land Prices in Seoul, South Korea / K.Jungsun, W. Jaewoong, K. Hyeongsoon, H. Joonghyeok. // Sustainability. – 2021. – №13(23). [in English]
3. Yunjong K. Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices / K. Yunjong, C. Seungwoo, Y. Mun Yong. // Sustainability. – 2020. – №12(14). [in English]

4. Crosby N. Property valuation variation and the 'margin of error' in the UK / N. Crosby, A. Lavers, J. Murdoch. // *Journal of Property Research*. – 1998. – №15. – p. 305–330. [in English]
5. Watkins C. The definition and identification of housing submarkets / Craig Watkins. // *Environment and Planning*. – 2001. – №33. – p. 2235 – 2253. [in English]
6. Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review / A.Craverro, S. Pardo, S. Sepulveda, L. Munoz. // 2022. – №12. [in English]
7. Evaluate your ML.NET model with metrics [Електронний ресурс] // microsoft.com. – 2023. – Режим доступу до ресурсу: <https://learn.microsoft.com/en-us/dotnet/machine-learning/resources/metrics>. [in English]
8. A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems / [J. Alfaro-Navarro, E. Cano, E. Alfaro-Cortes та ін.]. // *Complexity*. – 2020. – №2020. – p. 1–12. [in English]
9. Quang T. Housing Price Prediction via Improved Machine Learning Techniques / T. Quang, N. Minh, D. Hy. // *Procedia Computer Science*. – 2020. – №174. – p. 433–442. [in English]
10. What Is Comparative Market Analysis (CMA) in Real Estate? [Digital resource] // Investopedia. – 2022. – Resource access: <https://www.investopedia.com/terms/c/comparative-market-analysis.asp>.