УДК 004.032.26 **DOI 10.33042/2522-1809-2022-4-171-2-6**

О.В. Русова[1], В.М. Бредіхін[1], В.І. Вербицька[2]

[1]*Харківський національний університет міського господарства імені О.М. Бекетова, Україна*
[2]*Харківський національний автомобільно-дорожній університет, Україна*

ВДОСКОНАЛЕННЯ ОЦІНКИ НЕРУХОМОСТІ ЗА ДОПОМОГОЮ СУЧАСНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

У статті розглядається завдання оцінки вартості житла в містах України. Представлено результати роботи програмного додатку який складається з двох частин. Перша програма виконує збір необхідних даних для навчання нейронної мережі, а друга надає інструменти для попереднього аналізу зібраних даних та навчання на їхній основі багатошарової нейронної мережі прямого поширення.

Ключові слова: *нейронні мережі, глибоке навчання, машинне навчання, регресія, прогнозування, оцінка, аналіз даних.*

Постановка проблеми

Кінцева оцінка вартості житла є досить приблизною. Незважаючи на те, що дана операція може вимагати значних витрат часу, а точність такої оцінки не завжди перебуває на високому рівні, переважна більшість покупців здійснює аналіз цін на обраний об'єкт самостійно. Тому компанії, які працюють в сфері продажу нерухомого майна, пропонують своїм клієнтам власну оцінку об'єктів, а приватні оцінювачі привносять у роботу свій власний досвід та навички. Послідовність виконання оцінки є досить неточною, і дослідження, проведені у Великобританії та Австралії, показали, що оцінки двох професіоналів можуть відрізнятися до 40% [1]. Вочевидь, добре навчена машина, на противагу людині, має змогу виконувати подібні завдання з більшою послідовністю та точністю. Для вирішення цих проблем використовують нейронні мережі, які передбачають контроль певного безперервного значення, як правило, ціни або якої-небудь ймовірності.

У ряді країн уже давно існує практика застосування штучних нейронних мереж для масової оцінки об'єктів нерухомості з метою вирахування податку на майно. Але прогнозування ціни на нерухомість є корисним і для особистого використання, тому що процес купівлі-продажу житла є складним завданням, що вимагає розгляду безлічі різних факторів, таких, як зміна попиту на квартири, зміну вартості квартир залежно від сезону і т. д. Для спрощення цього завдання ми вважаємо доцільним використовувати нейронні мережі, які ґрунтуються на безлічі даних щодо поточного стану ринку нерухомості.

Аналіз останніх досліджень і публікацій

Компанії з країн СНГ та іноземні компанії (наприклад, SAP, Microsoft, IBM та ін.) надають як готові інструменти для клієнтської аналітики, так і розробляють індивідуальні рішення, налаштовані під окремих замовників. Найбільші банки, телеком-оператори, ритейлери мають у своєму арсеналі інструменти аналізу великих масивів даних і розробляють власні рішення щодо аналізу та прогнозу відтоку клієнтів (МТС, Ощадбанк, ПриватБанк).

Як правило, існуючі розробки ґрунтуються на особистісних даних клієнта, а також на даних про його активність у компанії: послуги та продукти, якими він користується, історію транзакційної активності, історію звернень, інформацію про покупки тощо. Отримані дані представляють собою великі масиви зі структурованою та неструктурованою інформацією, для аналізу та виявлення прихованих закономірностей яких широко використовується інтелектуальний аналіз даних та засновані на ньому методи машинного навчання [2].

Мета статті

Метою даної роботи є вдосконалення процесу аналізу даних ринку нерухомості за допомогою технологій машинного навчання.

Виклад основного матеріалу дослідження

Найчастіше ріелторські компанії надають платні послуги аналізу ринку нерухомості й оцінки вартості квартири. В той самий час в інтернеті наявні й безкоштовні сервіси оцінки квартир, такі як «LIGA.net» [3], ФДМУ [4] що здійснюють оцінку за

допомогою нейронних мереж. Негативним фактором даних сервісів (у порівнянні з додатком власної розробки для аналізу квартир у окремому місті) є те, що вони використовують єдиний загальний алгоритм для всіх міст і регіонів, не враховуючи розташування об'єктів нерухомості щодо місцевої інфраструктури, а також влучення у вибірку неправильних і помилкових оголошень, що негативно впливають на оцінку вартості квартир (наприклад, сайт не враховує, коли селище Перемога плутають із площею Перемоги). Проведений у даній роботі попередній аналіз зібраних даних щодо квартир, які присутні на ринку продажу нерухомості, дозволяє виправити ці помилки.

Отже, внаслідок дії великого обсягу показників і набору даних, необхідних для побудови моделі прогнозування, в якості механізму реалізації обчислень і візуалізації даних доцільно використовувати інформаційні технології. Різновидом означених інформаційних сучасних засобів може виступити спеціальне середовище розробки програм прогнозування мовою програмування Python.

Сучасне глибоке навчання пропонує розвинену інфраструктуру навчання нейронних мереж із вчителем. Завдяки додаванню нових шарів і блоків у межах одного шару глибока нейронна мережа може реалізовувати все більш складні функції.

Глибокі мережі прямого поширення, які також називають багатошаровими персептронами – найпоширеніші приклади моделей глибокого навчання. Метою такої мережі є апроксимація деякої функції f*. Глибока мережа прямого поширення визначає відображення $y = f(x, w)$ і за допомогою навчання знаходить параметри w, що дають найточнішу апроксимацію [5].

Для визначення вихідного значення нейрона, залежно від результату зваженої суми та граничного значення, використовуються функції активації. У теперішній час найпоширенішою функцією активації є *Relu*, яка дорівнює 0, якщо аргумент негативний, і дорівнює аргументу, якщо аргумент позитивний.

Основною перевагою функції *Relu* є простота її обчислення, що дуже важливо при використанні великих мереж даних. У даній роботі також було використано функцію активації *Relu* [6].

Завдання навчання глибокої мережі прямого поширення зводиться до завдання безумовної оптимізації. Таким чином, необхідно знайти точку мінімуму функції $E(w)$, *де* $w = [w_1, w_2, \ldots, w_n]T$ – вектор ваг мережі. Для вирішення даного завдання необхідно знайти такий вектор ваг w, на якому функція помилки мережі $E(w)$ прийме найменше значення. Найчастіше для цього використовуються градієнтні методи першого порядку, наприклад, метод градієнтного спуску. У даній роботі для вирішення завдання оптимізації використовується алгоритм *adaptive moments (Adam*), у якому для різних параметрів застосовуються різні швидкості навчання [7].

Для вирішення завдання оцінки вартості житла в Харкові було розроблено дві програми мовою програмування *Python*. Перша програма збирає необхідні для навчання нейронної мережі дані про квартири з оголошень сайту *OLX* [8], після чого структурує їх і записує в *csv*-файл. Друга програма надає інструменти для попереднього аналізу зібраних даних, після чого відбувається їхнє очищення, розподіл на навчальну та тестову вибірки і навчання на їхній основі багатошарової нейронної мережі прямого поширення.

Для збору, структурування та запису даних з сайту *OLX* були використані бібліотеки Requests і Beautiful Soup, написані мовою Python. Вибір обґрунтовується тим, що дані бібліотеки являють собою високорівневі інструменти, які дозволяють без додаткових налаштувань зробити http-запит і представити отримані дані у вигляді дерева синтаксичного аналізу, який дає можливості зручного отримання необхідних даних.

Програмою для збору даних було проведено аналіз кожного оголошення: у вихідному коді сторінки по відповідних до тегів даних зберігалися в *csv*-файл такі параметри, як номер оголошення, район і адреса квартири, вид об'єкта, поверх розташування даної квартири, загальна кількість поверхів у будинку, тип матеріалу будинку, кількість житлових кімнат у квартирі, загальна та житлова площі квартири, площа кухні та, відповідно, пропонована продавцем ціна. Також з *javascript*-модуля карти, що перебуває в кожному оголошенні, були виявлені і збережені у тому ж csv-файлі координати квартири для подальшої перевірки на віддаленість від центру міста та інших об'єктів інфраструктури.

В табл. 1 представлено фрагмент у форматі отриманих даних, зібраних першою програмою в *csv*-файл.

У другій програмі за допомогою інструментів *pandas* і *numpy* з отриманих даних були вилучені всі співпадаючі приклади (дублікати), оголошення з невірно зазначеними параметрами (наприклад, у графі «кількість кімнат»), а також усі оголошення, по координатах міста, що не попадають у межі міста Харків, що випадково потрапили у вибірку.

Таблиця 1

Перші шість записів у вихідному csv-файлі

Тип продажу	Поверх	Поверхів	Матеріал	Кімнат	Загальна площа	Житлова площа	Площа кухні	Район	X	У	Ціна
Вторинне житло	19	22	монолітний	2	52.3	26	16.5	Шевченківський	50.190	36.062	4500000
Вторинне житло	9	12	цегельний	2	41.6	19	5	Київський	50.183	36.069	2330000
Новобудова	3	18	цегельний	2	80.2			Київський	50.229	36.935	4251130
Новобудова	2	17	монолітний	1	39.1			Новобаварський	50.182	36.056	2112480
Вторинне житло	7	9	цегельний	1	28.5		6.5	Салтівський	50.191	36.015	2150000
Вторинне житло	7	10	панельний	1	34	18	7	Салтівський	50.147	36.028	1700000

Далі було проведено нормалізацію даних, тобто приведення до виду, здатного бути сприйнятим програмним забезпеченням. Дані типу «Object» наведені до числового типу.

Вибір потрібних характеристик значно впливає на підсумковий результат, тому займає, як правило, більше часу, ніж саме навчання моделі. Первинна оцінка характеристик проведена на основі кореляційної матриці. Коефіцієнти кореляції визначені за допомогою методу кореляції Пірсона (рис. 1).

Кореляційна матриця:
price
price 1,000000
area_value 0,928225
floors_total 0,315381
location_latitude –0,153215
location_longitude –0,209891
floor 0,238655
rooms 0,847558
…
building_type –0,017732
building_series –0,008270

Рис. 1. Кореляційна матриця характеристик нерухомого майна

З 12 характеристик на формування ціни впливають 10. Інші 2 характеристики мають оцінку кореляції менше 0,001 у порівнянні із ціною (price), тому не використовуються в дослідженні.

Після аналізу основних статистичних характеристик даних по кожній з числових ознак було виявлено, що всього лише 22 оголошення містять інформацію про продаж квартир з кімнатами більше чотирьох. Щоб уникнути дисбалансу класів дані приклади були вилучені. Після видалення всіх зайвих прикладів у наборі залишилося 3257 записів.

Наступний крок передбачав заповнення пропусків у даних. Пропуски були наявні в стовпцях «житлова площа» і «площа кухні». При видаленні рядків із пропущеними значеннями була присутня можливість втрати багатьох даних, необхідних для дослідження, що знизило б репрезентативність навчальної і тестової безлічей. Тому було вирішено здійснити заповнення цих даних методом підстановки з добором усередині груп. Пропущені дані в стовпцях «житлова площа» і «площа кухні» були заповнені середніми значеннями для квартир з відповідною до пропущених рядків кількістю кімнат.

Для роботи з нейронною мережею всі необхідні дані, такі як район, тип продажу і матеріал будинку, були наведені до категоріального виду за допомогою функції *categorical* інструмента *pandas.*

Для підвищення точності прогнозованої ціни доданий параметр «відстань від центру», що містить евклідову відстань точки знаходження квартири від центру міста, заданого широтою і довготою розташування обласного Національного банку України міста Харкова (49,995º північної широти і 36,233º східної довготи).

Таким чином був отриманий набір даних, що містить 22 стовпчика ознак і 3257 рядка прикладів. Для навчання нейронної мережі від даного набору відділяється цільова змінна Y («ціна»), після чого отримані набори розподіляться на навчальну і тестову вибірки в співвідношенні 80% до 20%. Після поділу навчальний набір *X_train* одержав 2605 прикладів для 21 стовпця. Тестовий набір *X_test* одержав 652 прикладів для 21 стовпця. Набори *y_train* і *y_test* з єдиним цільовим стовпцем «ціна» одержали по 2605 і 652 прикладів відповідно.

Як було сказано раніше, архітектура мережі заснована на глибокій нейронній мережі прямого поширення. Дана мережа реалізується повнозв'язною моделлю прямого поширення Sequential бібліотеки keras. Keras являє собою відкриту нейромережеву бібліотеку, написану мовою *Python*. Вибір бібліотеки обумовлений тим, що keras є простим і зручним інструментом у порівнянні з іншими, більш низькорівневими бібліотеками, такими як tensorflow, наприклад [9].

Для проектування нейронної мережі використовується layers API бібліотеки keras, що дозволяє користувачу створювати довільні шари. Для регуляризації застосовується інструмент keras.regularizers, що також перебуває в layers API. Для настроювання метрик моделі використовується метод compile. Було визначено три сховані шари, для кожного з яких було введено 512 нейронів і обрана функція активації *Relu.* У якості алгоритму оптимізації був обраний *Adam*, у якості метрик – середньоквадратична та середня абсолютна помилки. Модель була навчена за фіксовану кількість епох, яка дорівнювала 100.

У результаті тестування навченої нейронної мережі на тестовому наборі, яка складалась з 652 прикладів, отримана середня абсолютна помилка 3570,88. Це означає, що в середньому модель при виконанні оцінки вартості квартири помиляється на 3570 грн. При зіставленні даної величини помилки із середньою вартістю квартири в 238 тис. грн. можна приблизно оцінити точність оцінки в 85 %, що є гарним результатом.

В результаті виконання програми були побудовані графіки середньоквадратичної помилки та середньої абсолютної помилки (рис. 2, 3). Слід відзначити, що на етапі навчання величина помилки зменшується, це означає, що не був виявлений ефект перенавчання і модель нейронної мережі навчалася добре.

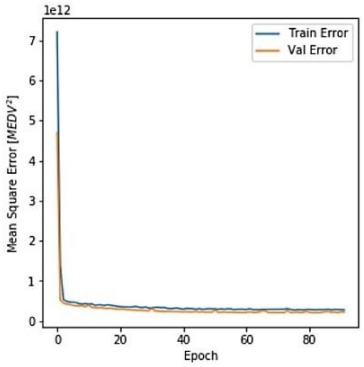

Рис. 2. Середньоквадратична помилка під час навчання

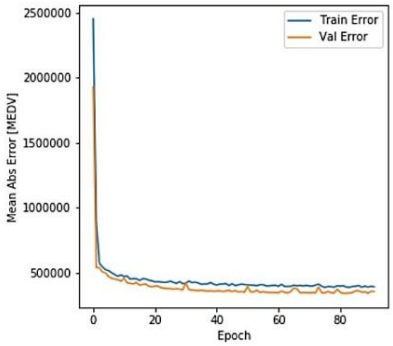

Рис. 3. Середня абсолютна помилка під час навчання

Таким чином, нейронна мережа досягла прийнятного рівня точності для оцінки вартості квартир у м. Харків. Були розроблені функції, що надають користувачу можливість за допомогою раніше навченої нейронної мережі оцінити вартість квартири за введеним даними.

Висновки

Таким чином, у рамках дослідження проведено аналіз факторів, що впливають на вартість оцінки нерухомості, в результаті чого виявлено потреба у використанні алгоритму машинного навчання, необхідного для аналізу цін на нерухомість. Для досягнення поставленої мети проведений кореляційний аналіз характеристик об'єктів нерухомості та ціни, що виявляють позитивний вплив на адекватність моделі. Здійснена попередня обробка даних шляхом видалення викидів повторюваних і незаповнених даних. Проведено ряд експериментів по добору гіперпараметрів алгоритму. Досягнуті характеристики та параметри дозволяють зробити висновок відносно адекватності моделі.

Були розроблені функції, що надають користувачу можливість за допомогою раніше навченої нейронної мережі оцінити вартість квартири за введеними даними. На окремих прикладах точність сягає 97%, що є позитивним фактом.

Література

1. Економічна наука та українські забудовники поки що шукають точки для перетину? [Електронний ресурс]. – Режим доступу: https://zn.ua/EONOMICS/zastrojshchiki-pytajutsja-prodavat-dovoennye-novostrojki-po-voennym-tsenam-prodazhi-stojat-tseny-na-ijul.html (дата звернення 15.08.2022)

2. Шаркаді М.М. Моделі і методи машинного навчання для завдань передбачення. [Текст] / М.М. Шаркаді, М.В. Роботишин, М.М. Маляр // Наук. вісник Ужгород. ун-ту. – 2020. - Вип. 36, № 1 - С. 112-121.

3. Ми оцінили вартість квартири через онлайн-сервіс ФДМ [Електронний ресурс]. – Режим доступу: https://biz.liga.net/all/nedvizhimost/ article/my-otsenili-stoimost-kvartiry-cherez-onlayn-servis-fgi-besplatno-kak-eto-rabotaet (дата звернення 17.08.2022)

4. Безкоштовний сервіс оцінки об'єктів нерухомості від Фонду державного майна України [Електронний ресурс]. – Режим доступу: https://evaluation.spfu.gov.ua. (дата звернення 17.08.2022)

5. Білашенко С.В. Розпізнавання зображень за допомогою згорткових нейронних мереж з використанням бібліотеки keras. [Текст] / С.В. Білашенко, Н.Н. Шаповалова, О.Г. Рибальченко // Гірничий вісник. - Вип. 103. - 2018. - С. 148-154 DOI: https://doi.org//10.31721/2306-5435-2018-1-103-153-158

6. Chaity Banerjee, Tathagata Mukherjee, Eduardo Pasiliao (2020) The Multi-phase ReLU Activation Function. ACM SE 20: Proceedings of the 2020 ACM Southeast Conference, 239–242. DOI: https://doi.org//10.1145/3374135.3385313

7. Rudenko O., Bezsonov O., Romanyk O., Lebediev V. (2019) Analysis of convergence of adaptive single-step

algorithms for the identification of non-stationary objects. Eastern-European Journal of Enterprise Technologies, 1(4), 6-14 DOI: https://doi.org//10.15587/1729-4061.2019.157288
8. Сайт OLX [Електронний ресурс]. – Режим доступу: https://olx.ua (дата звернення 30.08.2022)
9. Tensorflow [Електронний ресурс]. – Режим доступу: https://tensorflow.org/about/bib (дата звернення 30.08.2022)

References

1. Ekonomycheskaia nauka i ukraynskye zastroishchyky poka tolko yshchut tochky dlia peresechenyia? URL: https://zn.ua/ECONOMICS/zastrojshchiki-pytajutsja-prodavat-dovoennye-novostrojki-po-voennym-tsenam-prodazhi-stojat-tseny-na-ijul.html (data zvernennia 15.08.2022)
2. Sharkadi M.M., Robotyshyn M.V., Maliar M.M. (2020) Modeli i metody mashynnoho navchannia dlia zavdan peredbachennia. *Nauk. visnyk Uzhhorod. un-tu, 36, 1,* 112-121.
3. My otsenyly stoymost kvartyr cherez onlain-servys FHY URL: https://biz.liga.net/all/nedvizhimost/ article/my-otsenili-stoimost-kvartiry-cherez-onlayn-servis-fgi-besplatno-kak-eto-rabotaet (data zvernennia 17.08.2022)
4. Bezkoshtovnyi servis otsinky obiektiv nerukhomosti vid Fondu derzhavnoho maina Ukrainy URL: https://evaluation.spfu.gov.ua. (data zvernennia 17.08.2022)
5. Bilashenko S.V., Shapovalova N.N., Rybalchenko O.G. (2018) Rozpiznavannia zobrazhen za dopomohoiu zghortkovykh neironnykh merezh z vykorystanniam biblioteky keras. *Hirnychyi visnyk, 103*, 148-154. DOI: https://doi.org//10.31721/2306-5435-2018-1-103-153-158
6. Chaity Banerjee, Tathagata Mukherjee, Eduardo Pasiliao (2020) The Multi-phase ReLU Activation Function. *ACM SE 20: Proceedings of the 2020 ACM Southeast Conference*, 239–242. DOI: https://doi.org//10.1145/3374135.3385313
7. Rudenko O., Bezsonov O., Romanyk O., Lebediev V. (2019) Analysis of convergence of adaptive single-step algorithms for the identification of non-stationary objects. *Eastern-European Journal of Enterprise Technologies, 1(4),* 6-14. DOI: https://doi.org//10.15587/1729-4061.2019.157288
8. Sait OLX URL: https://olx.ua (data zvernennia 30.08.2022)
9. Tensorflow URL: https://tensorflow.org/about/bib (data zvernennia 30.08.2022)

Рецензент: д-р техн. наук, проф. А.Л. Литвинов, Харківський національний університет міського господарства імені О.М. Бекетова, Україна.

Автор: РУСОВА Ольга Віталіївна
студентка 2 курсу магістратури навчально-наукового інституту енергетичної, інформаційної та транспортної інфраструктури
Харківський національний університет міського господарства імені О.М. Бекетова
E-mail – olga.rusova@kname.edu.ua

Автор: БРЕДІХІН Володимир Михайлович
кандидат технічних наук, доцент, доцент кафедри комп'ютерних наук та інформаційних технологій
Харківський національний університет міського господарства імені О.М. Бекетова
E-mail – bredixinv@gmail.com
ID ORCID: https://orcid.org/0000-0002-6063-5046

Автор: ВЕРБИЦЬКА Вікторія Іванівна
кандидат економічних наук, доцент, доцент каф. обліку і оподаткування
Харківський національний автомобільно-дорожній університет
E-mail – verbytska67@gmail.com
ID ORCID: https://orcid.org/0000-0001-7103-6738

FOCUS ON ARTIFICIAL INTELLIGENCE FOR PREDICTING THE OUTFLOW OF CLIENTS FROM ON-LINE EDUCATION SITES

O. Rusova[1], V. Bredikhin [1], V. Verbytska [2]
[1] O. M. Beketov National University of Urban Economy in Kharkiv, Ukraine
[2] Kharkiv National Automobile and Highway University, Ukraine

The article examines the task of assessing the cost of housing in the cities of Ukraine. The purpose of this work is to simplify the determination of the value of apartments on the real estate market using machine learning technologies. To solve this problem, it is proposed to use a program module in Python using the Sequential direct distribution model of the keras library. A program was created that estimates the value of apartments according to their parameters using a neural network. The importance of forecasting in the field of real estate is shown, because the housing market is a systemic part of the regional economy. The results of the software application, which consists of two parts, are presented. The first program collects the necessary data for training a neural network about apartments from the OLX site ads, their structuring and recording in a csv file. The second program provides tools for preliminary analysis of the collected data, after which they are cleaned, divided into training and test samples and trained on their basis by a multilayer neural network of direct propagation using a machine learning algorithm. The layers API of the keras library was used to design the neural network, which allows the user to create arbitrary layers. For regularization, the keras.regularizers tool, which is also in the layers API, is used. To configure model metrics, the compile method was used. Three hidden layers were defined, for each of which 512 neurons were introduced and the Relu activation function was chosen. Calculations of the correlation of prediction indicators and error curves of machine learning are given. As a result of testing the trained neural network on a test set of 652 examples, an average absolute error of 3570.88 was obtained, and the accuracy of the model was approximately 85%. Thus, the neural network has reached an acceptable level of accuracy for estimating the cost of apartments in the city of Kharkiv. Ways to reduce test errors and learning errors using cross-validation are proposed. Concepts of learning hyper-parameters and their regularization are considered

Keywords: *neural networks, deep learning, machine learning, regression, prediction, estimation, data analysis.*