УДК 004.89

Візнюк А.В. студент гр. ІПЗ-42
(*Київський національний університет імені Тараса Шевченка, м. Київ, Україна*)
Науковий керівник: Юрчук І.А., к.ф.-м.н., доцент кафедри програмних систем і технологій
(*Київський національний університет імені Тараса Шевченка, м. Київ, Україна*)
Дяченко Г.Г., асистент кафедри електропривода
(Національний технічний університет "Дніпровська політехніка", м. Дніпро, Україна)

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ ВСТАНОВЛЕННЯ ВАРТОСТІ ВЖИВАНИХ АВТОМОБІЛІВ ЗГІДНО ЇХ АТРИБУТІВ

Актуальність. На придбання авто може впливати безліч факторів – модель, рік випуску, потужність двигуна, пробіг тощо. В той же час може виникнути необхідність встановити справедливу ціну на авто для продажу. Запропонований програмний застосунок аналізує зібрані дані з інших транспортних засобів з метою встановлення ціни на автомобіль в залежності від значення його атрибутів.

Збір даних. Дані для створення моделі машинного навчання були зібрані з веб-ресурсів вторинного ринку України. Кількість записів налічує 44259 рядків. До уваги були взяті наступні атрибути автомобілів: виробник, модель, тип кузова, тип приводу, тип трансмісії, тип палива, двигун, об’єм двигуна (в літрах), кінські сили, пробіг (в км.), рік випуску та ціна (в доларах).

Модель машинного навчання. В якості моделі для машинного навчання було вирішено обрати нейронну мережу зі зворотнім поширенням похибки (Backpropagation network), оскільки нейронні мережі є універсальними апроксиматорами і, на відміну від лінійної регресії, можуть знаходити нелінійні залежності між характеристиками [1].

Оскільки нейронна мережа може працювати лише з числовими даними, то категоріальні характеристики виробник, модель, тип кузова, тип приводу, тип трансмісії та тип палива потрібно закодувати в числовий формат. Оскільки зазначені категорії не мають логічного порядку розташування за значимістю, то стандартним вибором кодування буде One Hot Encoding [2], в якому замінюємо кожну категорію числовим вектором, що складається з нулів та одиниць. Проблеми, які можуть виникнути при застосуванні даного кодування пов’язані з тим, що для числа n унікальних категорій, розмірність числового вектору буде також n. Чим більше унікальних категорій – тим більше вхідних значень до нейронної мережі. Для зменшення кількості категорій було вирішено об’єднати ті, що зустрічаються в наборі даних з частотою меншою за 0.1%. Таким чином, число унікальних категорій було зменшено з 1663 до 276.

Перед початком навчання нейронної мережі числові характеристики, такі як: об’єм двигуна, кінські сили, пробіг, рік випуску, потрібно нормалізувати / стандартизувати, оскільки вони позначають різні фізичні величини та можуть варіюватися в різних діапазонах. Це може спричинити більший вплив однієї з величин на результат роботи моделі. Кожна числова характеристика була стандартизована за наступним співвідношенням: $x' = (x - \bar{x})\sigma^{-1}$ [3].

Розроблена нейронна мережа має наступну структуру: 280 вхідних значень, 512, 256, 128 нейронів у внутрішніх шарах відповідно і 1 нейрон на виході. Активаційні функції в внутрішніх шарах – $max(x, 0)$ (rectified linear unit, ReLU), що є стандартним вибором активаційної функції в нейронних мережах на сьогодні [4]. Функція втрат – середня квадратична помилка (Mean Squared Error, MSE).

Враховуючи велику кількість вхідних даних (44259 записів), в якості алгоритму для навчання нейронної мережі було надано перевагу міні-пакетному градієнтному спуску

(Mini Batch Gradient Descent, MBGD). Він поєднує переваги стохастичного градієнтного спуску (Stochastic Gradient Descent, SGD) та пакетного (Batch Gradient Descent, BGD), обчислюючи середню помилку не для одного значення з вибірки чи з всього набору даних, а з групи зразків наперед заданої довжини. Це пришвидшує процес навчання та зменшує дисперсію оновлень параметрів, що призводить до стабільної збіжності [5].

Аналіз результатів. Для визначення якості розробленої моделі були використані наступні метрики (кількість епох навчання – 200, розмір пакету batch – 64):

- Середня абсолютна похибка (Mean Absolute Error, MAE) становила 1659.
- Коефіцієнт детермінації (R^2) становив 0,88.

Для перших 100 рядків даних з тестової вибірки було обчислено значення ціни за допомогою нейронної мережі та порівняно з актуальними цінами (див. рис. 1):

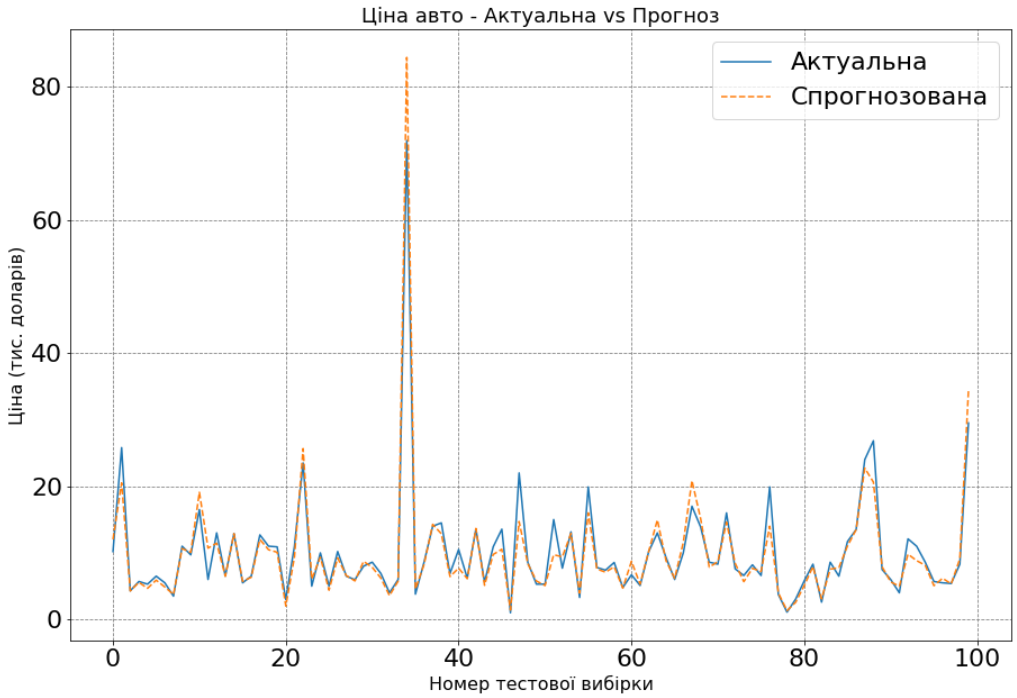

Рисунок 1 – графік порівняння справжньої ціни та ціни, що була встановлена нейронною мережею

Висновки та перспективи подальшого вдосконалення. Результати роботи нейронної мережі показують, що встановленні значення ціни досить сильно корелюють з актуальними значеннями, про що свідчить високий коефіцієнт детермінації – 0,88. Подальша робота передбачає застосування моделей машинного навчання, таких як: дерево прийняття рішень, лінійна регресія, випадковий ліс і т. д. для встановлення ціни авто в залежності від значення атрибутів.

Перелік посилань

1. Порівняльний аналіз моделей машинного навчання і регресій для прогнозування ціни легкового авто. Вісник Харківського національного університету імені В. Н. Каразіна серія «Економічна». 2019. No. 97. С. 31–40.

2. Seger, C. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. Degree project technology. 2018.

3. How to use Data Scaling Improve Deep Learning Model Stability and Performance: URL: https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/ (дата звернення: 04.11.21).

4. Activation Functions in Neural Networks | by SAGAR SHARMA | Towards Data Science: URL: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6 (дата звернення: 04.11.21).

5. Batch, Mini Batch & Stochastic Gradient Descent | by Sushant Patrikar | Towards Data Science: URL: https://towardsdatascience.com/batch-mini-batch-stochastic-gradient-descent-7a62ecba642a (дата звернення: 04.11.21).