DOI 10.31891/2307-5732-2023-321-3-83-86
УДК 621

ПАВЛІЧКО ВЛАДИСЛАВ
Національний університет "Львівська політехніка"
ORCID ID: 0000-0001-9596-5240
e-mail: vladyslav.pavlichko@gmail.com
МЕЛЬНИКОВА НАТАЛІЯ
Національний університет "Львівська політехніка"
ORCID ID: 0000-0002-2114-3436
e-mail: melnykovanatalia@gmail.com

ПЕРЕДБАЧЕННЯ ЦІНИ АВТОМОБІЛЯ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Розвиток сучасного машинобудування є наслідком швидкого зростання економіки в багатьох розвинутих країнах. Станом на сьогоднішній день, інтеграцію автомобілів у повсякденне життя складно заперечити. Логістика, громадський транспорт, особистий транспорт, сервіси доставки, таксопарки, оренда авто тощо. Відповідно, маючи таку гнучкість використання, з'являється і попит на дану продукцію. В свою чергу, виробники прагнуть отримати найбільшу кількість покупців, що призводить до утворення конкуренції на ринку. Як результат час, який витрачається на підбір бажаного авто, є суттєвим, і сам процес вивчення ринку і характеристик всіх авто не є доцільним для користувачів. Тому це дослідження буде спрямовано на спрощення цього процесу, шляхом використання методів машинного навчання, що дасть змогу швидше зорієнтуватись у цінах на ринку і швидше обирати бажане авто.

В роботі наведено результати досліджень методів машинного навчання щодо передбачення ціни автомобіля, попередньої обробки тренувальних даних; запропоновано власну реалізацію, яка базується на комбінуванні декількох моделей машинного навчання. За основу для досліджень були взяті дерева рішень і випадковий ліс, обговорено загальну концепцію їхньої роботи, переваги і недоліки, алгоритм їхньої побудови. Метою цієї статті є порівняльна характеристика застосованих моделей. Окремі моделі можуть прогнозувати результати з високою точністю: дерева рішень – 90% точність і випадковий ліс – 95%. Однак точність являється не єдиною метрикою оцінки ефективності роботи моделей. Впливовість MAE, RMSE є невід'ємними метриками для оптимальної роботи моделей. Згідно з результатами досліджень, значення обидвох метрик кращі у випадкового лісу, що доказує ефективність даного рішення. Однак результати можна покращити, використовуючи комбінований підхід. Стаття якраз і надаватиме відповідь на питання ефективності застосування комбінованих підходів.

Ключові слова: передбачення ціни автомобіля, машинне навчання, попередня обробка, машинобудування.

PAVLICHKO VLADYSLAV T., MELNYKOVA NATALIIA I.
Lviv Polytechnic National University

CAR PRICE PREDICTION USING ML METHODS

The development of car manufacturing industry is consequence of fast-growing economics in developed countries. It is hard to deny the fact of strong integration of cars in all spheres of human life nowadays. Logistics, public transport, personal usage, delivery services, taxi depots, cars for rent, etc. Demand increases due to its flexibility of usage, manufacturers strive to increase amount of customers, which creates competition on the market. As a result, the amount of time spent on investigation of market and key features of each car is dramatically huge. That is why the aim of this research is to simplify process of observation of cars, their features and prices, using machine learning techniques, which will allow to choose desirable cars in shorter time.

Article offers researches of applying machine learning methods for car price prediction, train data preprocessing, suggesting own approach based on combination of several machine learning models. Decision trees and random forest models were chosen as basic methods for this research, general concepts and construction algorithms are discussed, highlighted pros and cons each of them. The main purpose of article is the comparison of proposed methods. Distinct models are capable to predict results quite preciously – 90% accuracy for decision trees and 95% for random forest. But R^2 (accuracy) is not the only metrics used in model's effectiveness evaluation. RMSE and MAE have crucial influence on optimal work of the model. Relying on the results of this research, random forest gives better RMSE and MAE values comparing to decision trees, which proves the effectiveness of such approach. However, results might improve with usage of combined approach. The question about effectiveness of such approach will be answered during this research.

Keywords: car price prediction, machine learning, pre-processing, mechanical engineering,

Постановка проблеми

Цікавість до вирішення даної проблеми є актуальною, оскільки автомобільна тематика тісно взаємопов'язана зі всіма сферами життя людей. Особливо корисною це дослідження буде у сфері бізнесу і особистого використання авто. Попит на авто зростає, автомобілями цікавляться все більше і все частіше їй купляють для особистого використання, або в якості сімейного авто. Однак для людини, яка не орієнтується в ринку авто, варіативність модельного виробників і модельного ряду може значно ускладнити процес вибору.

Для випадку з бізнесом ситуація дещо змінюється. Метою будь-якого бізнесу є мінімізація витрат і максимізація чистого прибутку, що, в свою чергу, призводить до детального планування використання коштів. Взявши умовно починаючий або середній бізнес, головною частиною якого є автотранспорт, наприклад таксопарк, перевезення вантажів, пошта, перевезення пасажирів, швидкий підбір авто допоможе зекономити час, що дасть змогу направити його для вирішення інших проблем бізнесу.

Аналіз останніх джерел

Дослідження у статті [1] спрямовані на поєднання методів популяційного рою, градієнтного підйому та мережі BP нейронів. Процес обробки даних детально описаний у статті, зображено кінцеві результати, проте недоліком є недостатньо описані всі параметри, що ускладнює відтворення поставлених експериментів.

У статті [2] проводиться аналіз ефективності різних моделей машинного навчання, а саме: регресія, дерев рішення та нейронні мережі. Модель нейронної мережі показала найкращу точність серед всіх. Додатково, автори наводять список факторів, які були використані для прогнозування цін на автомобілі, що може бути корисним при проведенні подібних досліджень. Мінусом статті є мала вибірка даних, оскільки обраний датасет є досить малим, що може призвести до некоректної роботи моделей, особливо з новими вхідними даними.

Стаття [3] описує дослідження використання випадкового лісу для поставленої задачі. В ході дослідів використовувалась логістична регресія для порівняння. Згідно з результатами експерименту, випадковий ліс виявився більш ефективним у порівнянні з логістичною регресією. Проте в статті не надано детального опису використаних даних, а також не наведено достатньо великого обсягу експериментальних результатів для оцінки ефективності запропонованого методу в порівнянні з іншими методами, що були використані раніше.

Виклад основного матеріалу

Оскільки дана проблема є комплексною, її вирішення можна розбити на декілька підзадач:

- Пошук вибірки даних і її попередню обробка
- Оцінка вагомості кожної з ознак
- Підбір методів машинного навчання
- Налаштування методів та їх використання

Для дослідження був обраний датасет [4] у відкритому доступі, з понад 400 тисяч лотів авто американського ринку. Попередня обробка такої вибірки буде обов'язковою, оскільки кількість неінформативних записів може бути значною. У наборі даних 26 колонок, які несуть в собі ту чи іншу інформацію, однак не всі характеристики будуть важливими для ціноутворення. Даний етап можна виокремити у препроцесінг [5], оскільки проводиться видалення записів з відсутньою інформацією, видалення викидів даних, а також виробників, кількість авто яких є суттєво малою. Результат обробки наведений на рис. 1.

	price	year	manufacturer	model
31	15000	2013.0	ford	f-150 xlt
32	27990	2012.0	gmc	sierra 2500 hd extended cab
33	34590	2016.0	chevrolet	silverado 1500 double
34	35000	2019.0	toyota	tacoma
35	29990	2016.0	chevrolet	colorado extended cab
...	...	...	...	...
426863	25590	2017.0	null	Genesis G80 3.8 Sedan 4D
426869	13990	2016.0	null	Scion iM Hatchback 4D
426870	22990	2020.0	hyundai	sonata se sedan 4d
426874	33590	2018.0	lexus	gs 350 sedan 4d
426878	28990	2018.0	lexus	es 350 sedan 4d

121838 rows × 14 columns Open in new tab

Рис. 1. Датасет після препроцесінгу

Наступний крок – виокремлення лише вагомих ознак, які впливають на фінальну ціну авто. Такі характеристики як опис авто не є показником, оскільки аналізувати опис кожного авто – не є однією з задач цього дослідження, скоріше її вже можна класифікувати як проблему зі сфери NLP, щоб виокремлювати головні аспекти з опису, що є цілою окремою роботою. До цього списку також відносяться поля: 'VIN', 'posting_date', 'id', 'url', 'region', 'state', 'region_url', 'image_url', 'lat', 'long', 'description', 'county'. Вони не несуть ніякої інформації для користувача і ніяк не впливають на ціну авто.

Нижче наведено кореляційну матрицю, що визначає залежності між ознаками. Найбільш важливою для цього дослідження – кореляція ціни з іншими ознаками. З рис. 2 можна зробити висновок, що найбільш вагомі ознаки – це рік, пробіг, кількість циліндрів у моторі і вид КПП.

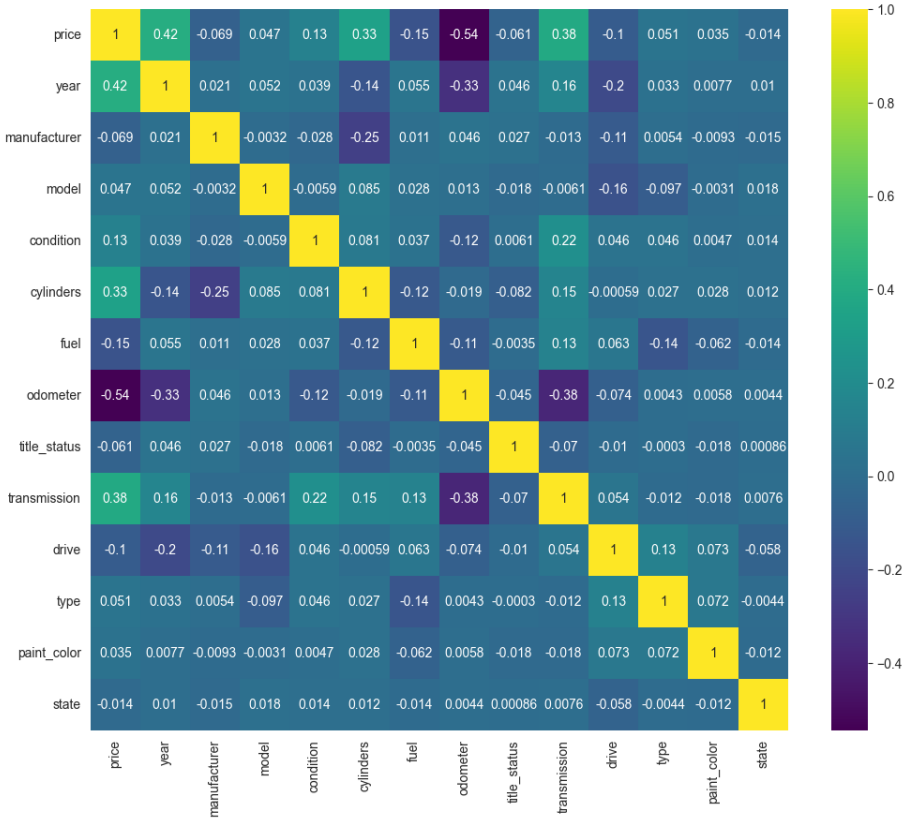

Рис. 2. Кореляційна матриця ознак

Методами, які використовуватимуться у дослідах було обрано дерева рішень (Decision Tree [6]) і випадковий ліс (RandomForest [7]). У проаналізованих роботах зазначалось, що результати, які досягаються з їхнім використанням є досить високими для точного передбачення, у порівнянні SVM, kNN та іншими.

Дерева рішень являють собою модель машинного навчання, які репрезентуються у вигляді дерев, переважно бінарного виду. Складаються вони з вузлів, ребер і листків [8]. Алгоритм роботи дерев [9] наступний:

- Головний вузол – точка вхідних даних.
- На основі вхідних даних, перехід з одного вузла у інший є результатам відповідності переліку правил у вузлі.
- Кожне ребро з'єднує вузли між собою, проте у випадку бінарної репрезентації, у вузла можуть бути лише 2 нащадки.
- Процедура відповідності правилам повторюється для інших ознак.
- Критерії зупинки – обмеження глибини дерева, рання зупинка, введення поняття мінімального допустимого.

Випадковий ліс – ансамблевий метод, який використовує в собі дерева рішень. Метод використовує в собі дерева рішень, проте з іншими цілями. При тренуванні, будується *n*-на кількість дерев з *m*-ознаками [10], обраними з вихідних даних. Як результат видується ансамбль дерев, які вчаться на різних ознака і на різних даних, які теж випадково беруться для кожних дерев. Таким чином вирішується проблема перенавчання, яка притаманна деревам рішень, а обирання прогнозу обирається мажоритарним голосуванням [11] або усередненням результатів зі всіх дерев.

Хоч використання одного методу для прогнозування ціни може надавати досить точні результати, проте для покращення нерідко комбінують методи між собою. У цій роботі буде запропоновано власний підхід, який базується на комбінації дерева рішень і випадкового лісу.

Реалізацію підходу можна поділити на два етапи: використання класифікаційного дерева рішень для формування додаткової вагомої ознаки – категорії, до якої відноситься авто і використання випадкового лісу, що зображено на рис. 3

З результатів помітно, що результати з використанням комбінування моделей є значно кращими, ніж використання моделей по окремості, про що свідчать значення R2, RMSE та MAE. Зменшення останніх двох метрик свідчить про те, що модель надає стабільніші результати, які суттєво не відрізняються від реальних цін. У конкретному випадку, у середньому, отримані спрогнозовані ціни варіюються близько 1187$, що не є великою різницею, особливо беручи до уваги, що ціни базуються на американському ринку.

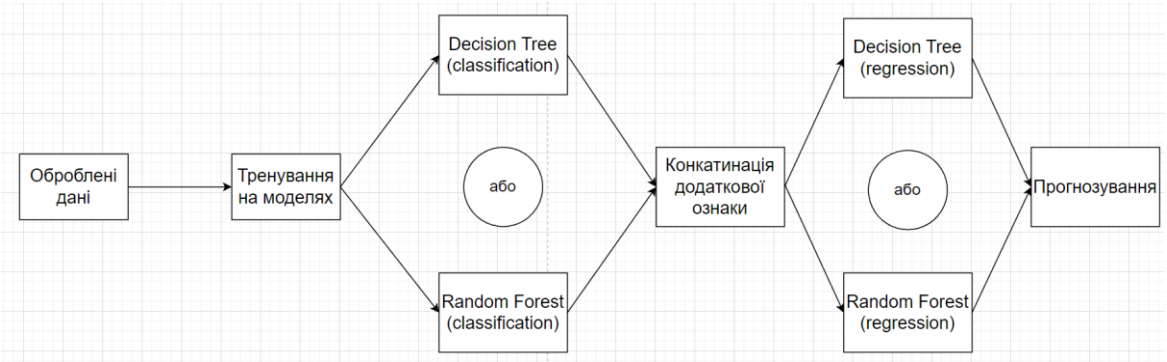

Рис. 3. Запропонований метод

Таблиця 1

Порівняння отриманих результатів

Модель	R2	RMSE	MAE
Decision Tree	0.90	3755.79	1820.45
Random Forest	0.95	2808.99	1550.68
Decision Tree +Random Forest	0.97	2061.64	1187.59

Висновки

У цій роботі було проведено досліджено тематику прогнозування ціни автомобіля. Результатами досліджень являється порівняльна таблиця застосованих підходів прогнозування. Застосовано наступні методи машинного навчання: дерева рішень, випадковий ліс і запропонований комбінований метод, який використовує для своєї роботи два попередніх. Одинарне використання методів хоч і надає точні результати, 90% і 95% для дерев рішень і випадкового лісу, проте середня абсолютна похибка залишається значною. Комбінований підхід, в свою чергу, дозволяє розбити процес прогнозування на декілька етапів і покращити кожну з використовуваних метрик – точність у 97% на тестових даних є солідним результатом. Середню абсолютну похибку близько 1000$ не можна назвати високою, оскільки бюджет для авто закладається з очікуванням підвищеної ціни.

За результатами досліджень, комбінування різних методів дозволяє покращити точність передбачення, зменшити помилку, що суттєво збільшить якість передбачень для кінцевого користувача. Подальший розвиток цього дослідження може включати застосування більшої кількості різних методів машинного навчання, тренування і застосування нейронних мереж і вдосконалення запропонованого підходу з метою покращення результатів.

References

1. Liu E., Li J., Zheng A. (2022). Research on the prediction model of the used car price in view of the pso-gra-bp neural network. Sustainability. Vol. 14, No. 15. P. 8993.

2. Gegic E., Isakovic B., Keco D. Car price prediction using machine learning techniques. Vol. 8, No. 1. P. 6.

3. Pandey A., Rastogi V., Singh S. Car's selling price prediction using random forest machine learning algorithm. Rochester, NY: 2020.

4. Used cars dataset. https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data.

5. Piramuthu S. Input data for decision trees. Expert Systems with Applications. 2008. Vol. 34, No. 2. P. 1220–1226.

6. Webb G. I., Fürnkranz J., Fürnkranz J., Sammut C. (2011). Decision tree. Encyclopedia of Machine Learning / G. I. Webb. Boston, MA: Springer US, P. 263–267.

7. Breiman L. Random forests. Machine Learning. 2001. Vol. 45, No. 1. P. 5–32.

8. Rokach L., Maimon O. Decision trees. The Data Mining and Knowledge Discovery Handbook. 2005. P. 165–192.

9. Garofalakis M., Hyun D., Rastogi R., Shim K. Building decision trees with constraints. Data Mining and Knowledge Discovery. 2003. Vol. 7, No. 2. P. 187–214.

10. Toloşi L., Lengauer T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. Vol. 27, No. 14. P. 1986–1994.

11. El Habib Daho M., N. Settouti, M. El Amine Lazouni, M. El Amine Chikh (2014). Weighted vote for trees aggregation in random forest.