УДК 004.8:338+629.33

В. В. Гіжевський
С. О. Жуков

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ЦІН НА ВЖИВАНІ АВТОМОБІЛІ

Вінницький національний технічний університет

Анотація

В роботі проведено аналіз предметної області передбачення цін на вживані автомобілі, попередньо запропоновано ознаки, які мають вплив на ціноутворення вживаних автомобілів. Здійснено огляд аналогічних рішень, запропоновано алгоритм створення ІТ передбачення цін на вживані автомобілі, на основі якого проведено створення ІТ. Виконано вибір та опис набору даних, проведено попереднє очищення даних. Проведено розвідувальний аналіз даних, запропоновано правила фільтрації аномальних значень, обрано моделі машинного навчання, здійснено їх тренування та визначено оптимальну модель.

Ключові слова: вживані автомобілі, інформаційні технології, машинне навчання, аналіз даних, передбачення ціни, ознаки, моделі машинного навчання, передбачення цін..

Abstract

The work analyzes the subject area of predicting prices for used cars and preliminarily proposes features that have an impact on the pricing of used cars. A review of similar solutions was carried out, an algorithm for creating IT for predicting prices for used cars was proposed, on the basis of which the IT was created. The data set is selected and described, and preliminary data cleaning is performed. An exploratory analysis of the data was carried out, rules for filtering anomalous values were proposed, machine learning models were selected, trained, and the optimal model was determined.

Keywords: used cars, information technology, machine learning, data analysis, price prediction, features, machine learning models, price prediction.

Вступ

Сучасний світ перебуває в постійному русі, де інформаційні технології стають ключовими для вирішення різних завдань і завдяки ним створюються нові можливості в багатьох сферах. Однією зі сфер, де інформаційні технології здійснюють значний вплив, є ринок вживаних автомобілів. Підвищення доступності інформації в інтернеті, а також застосування аналітичних і алгоритмічних методів у сфері економіки та фінансів відкривають перед дослідниками та бізнесменами нові перспективи для аналізу та передбачення цін на вживані автомобілі.

Актуальність теми дослідження полягає в кількох ключових аспектах:

1. Зростання ринку вживаних автомобілів: Ринок вживаних автомобілів постійно росте, як в розвинених, так і в країнах що розвиваються. Це робить його важливим об'єктом дослідження та управління.
2. Зміна споживацьких звичок: Сучасні споживачі все більше віддають перевагу покупці вживаних автомобілів, через їхню доступність і менші витрати. Знання про ціни та передбачення їхньої динаміки стають критичними для покупців та продавців.
3. Зростання даних та обчислювальної потужності: Споживання та зберігання даних про ринок вживаних автомобілів значно зросло, що робить можливим застосування сучасних аналітичних методів для зрозуміння ринкової динаміки та передбачення цінових тенденцій.
4. Фінансовий аспект: Ринок вживаних автомобілів є значущою галуззю у фінансовому сенсі, і правильне управління цінами може призвести до збільшення прибутку та оптимізації ресурсів.
5. Споживчі переваги: Дослідження цінових тенденцій на ринку вживаних автомобілів може допомогти споживачам знайти оптимальні пропозиції, зберігаючи їхні фінанси та задоволення потреб в транспорті.

Зважаючи на ці аспекти, дослідження в галузі інформаційних технологій для аналізу та передбачення цін на вживані автомобілі має важливе значення і може призвести до покращення ефективності на ринку та сприяти сталому розвитку автомобільної індустрії в цілому [1].

Результати дослідження

Для проведення дослідження було обрано набір даних, що має назву “Used Cars Dataset” та опублікований користувачем Austin Reese та доступний для загального використання на платформі Kaggle [2]. Цей датасет включає в себе широкий спектр інформації про продаж вживаних автомобілів (по 426880 автомобілям), яку надає компанія Craigslist. Зокрема, серед стовпців можна виділити такі параметри, як ціна, рік, стан автомобіля, виробник, координати (широта/довгота) та ще 20 інших категорій (рис. 1).

price	year	manufacturer	model	condition	cylinders	...	size	type	paint_color	image_url	description	county	state
33590	2014.0	gmc	sierra 1500 crew cab slt	good	8 cylinders	...	NaN	pickup	white	https://images.craigslist.org/00R0R_lwWjXSEWNa...	Carvana is the safer way to buy a car During t...	NaN	ca
16999	2014.0	chevrolet	express cargo	excellent	NaN	...	NaN	NaN	white	https://images.craigslist.org/00101_eEfFA8nmAr...	Ready To Upgrade Your Ride Today? We Make It F...	NaN	wa
4795	2007.0	acura	mdx	NaN	6 cylinders	...	NaN	SUV	NaN	https://images.craigslist.org/00M0M_1xUnC7COzh...	2007 *Acura* *MDX* 4WD 4dr Tech Pkg SUV - $4,7...	NaN	ga
21995	2018.0	toyota	c-hr	excellent	NaN	...	NaN	wagon	white	https://images.craigslist.org/00a0a_iCJnx7a1Sa...	2018 Toyota C-HR XLE 4dr Crossover Wagon Rea...	NaN	ca
27395	2015.0	ford	f-150	NaN	6 cylinders	...	full-size	truck	black	https://images.craigslist.org/00J0J_gjwKlmegPk...	2015 Ford F150 4x4 Truck Lariat Navigation Ext...	NaN	nc

Рис. 1. Приклад ознак автомобілів, що містить набір даних

Проведено попереднє очищення даних. Дослідивши інформацію по кожній ознаці датасету, виявлено, що є такі ознаки, що мають велику кількість пустих значень, або таких значень, що не несуть ніякої цінності для тренування моделі машинного навчання, тому, було прийнято рішення їх видалити а деякі заповнити значенням «unknown». В результаті очищення, датасет зменшився до 203829 автомобілів.

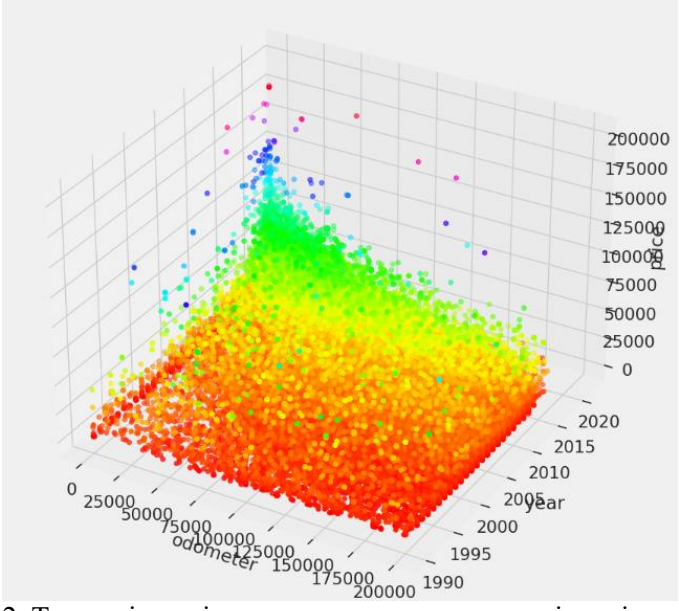

Рис. 2. Тривимірна діаграма даних за ознаками ціна, рік, пробіг

З рисунку 2 видно тривимірну діаграму розподілу автомобілів за ціною, роком випуску та пробігом. Більша частина автомобілів знаходиться після 2005 року. Сильний зріст кількості автомобілів видно з 2008 по 2020 роки. Також видно, як з роками, ціна на новіші автомобілі, зростає, досить різко. Також, в нижньому лівому куті діаграми, де пробіг автомобілів, що є досить старими, знаходиться між значенням нуля та 50 тисяч зустрічається рідше. Але в той же час, автомобілі, що мають недавній рік випуску, та малий пробіг – більшість на цій діаграмі.

Задача передбачення цін на вживані автомобілі відноситься до виду машинного навчання з вчителем. Визначено, що наша задача відноситься до задачі регресії, оскільки нам потрібно передбачувати значення змінної на основі набору ознак, але також, нам потрібні моделі які побудовані на основі дерев рішень [3].

Для визначення цін на вживані автомобілі необхідно створити, налаштувати та провести тренування моделей машинного навчання. Пропонується використовувати наступні моделі: Linear Regression, Support Vector Machines, Linear SVR, Stochastic Gradient Decent, Decision Tree Regressor, RandomForestRegressor, XGBoostRegressor, LGBM, GradientBoostingRegressor, RidgeRegressor, BaggingRegressor, ExtraTreesRegressor, AdaBoostRegressor, VotingRegressor.

Результат тренування моделей відображено в точності передбачення цільової ознаки (ціни) за трьома критеріями: коефіцієнтом детермінації R^2, відносною похибкою та середньоквадратичним відхиленням RMSE (рис.3) [4].

	Model	r2_test	d_test	rmse_test
7	LGBM	0.92	15.42	3,773.23
5	Random Forest	0.90	16.72	4,276.71
11	ExtraTreesRegressor	0.89	16.16	4,354.25
10	BaggingRegressor	0.89	17.61	4,502.19
6	XGB	0.88	21.50	4,720.92
4	Decision Tree Regressor	0.82	19.52	5,672.66
8	GradientBoostingRegressor	0.74	33.46	6,808.29
13	VotingRegressor	0.54	52.54	9,119.55
0	Linear Regression	0.54	52.46	9,118.34
9	RidgeRegressor	0.54	52.46	9,118.34
3	Stochastic Gradient Decent	0.54	52.42	9,127.74
2	Linear SVR	0.47	43.01	9,786.91
12	AdaBoostRegressor	0.41	86.17	10,327.89
1	Support Vector Machines	0.33	61.17	11,007.60

Рис. 3. Точність моделей за трьома критеріями

На рисунку 3 видно, що найкращою моделюю за усіма показниками є модель LGBM [5], її тренування дозволило отримати точність передбачення, за критерієм детермінації R^2, 0.92 (чим ближче значення до 1, тим краще точність).

Дане рішення дає кращий результат в точності, порівнюючи його з аналогами, що використовують подібні моделі. Серед яких:

- “XGBoostRegressor, R2_score = 0.896” [6];
- “RandomForestRegressor, R2_score = 0.867” [7].

Висновки

Під час дослідження набору даних “Used Cars Dataset”, що містить інформацію по продажу вживаних автомобілів на порталі Craigslist, проведено розгорнутий розвідувальний аналіз, фільтрування даних та очищення аномальних значень. Побудовано тривимірну діаграму, що показує розподіл автомобілів за трьома ознаками: ціна, рік випуску, пробіг автомобіля.

Наступним кроком стала побудова моделей для передбачення цільової ознаки, їх налагодження та тренування, порівняння останніх за допомогою критеріїв визначення точності, щоб визначити, які серед них є оптимально найкращою. Для дослідження було обрано регресійні моделі та моделі на основі дерев рішень.

Найкращий результат серед усіх використаних моделей дала модель LGBM. Її використання дозволило отримати точність передбачення 0.92 за критерієм детермінації, що є кращим результатом за найкращі аналоги із значенням R2 – 0.896, та менше.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. В. Б. Мокін, А. В. Лосенко, і М. В. Дратований, «ІНТЕЛЕКТУАЛЬНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ЦІН НА ВЖИВАНІ АВТОМОБІЛІ», *Вісник ВПІ*, вип. 6, с. 62–72, Груд. 2019.
2. Used Cars Dataset. Kaggle. 2021 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
3. The Python Data Science Handbook by Jake VanderPlas (O’Reilly). 2016 [Електронний ресурс] – Режим доступу: https://jakevdp.github.io/PythonDataScienceHandbook/
4. Interpretation of Evaluation Metrics For Regression Analysis (MAE, MSE, RMSE, MAPE, R-Squared, And Adjusted R-Squared). Medium. May 24, 2022 [Електронний ресурс]. – Режим доступу: https://medium.com/@ooemma83/interpretation-of-evaluation-metrics-for-regression-analysis-mae-mse-rmse-mape-r-squared-and-5693b61a9833
5. LightGBM documentation. Microsoft Corporation. 2023 [Електронний ресурс]. – Режим доступу: https://lightgbm.readthedocs.io/en/stable/
6. Used Car Price Prediction using Machine Learning. Piblished in Towards Data Science. Panwar Abhash Anil. Aug 3, 2020 [Електронний ресурс]. – Режим доступу: https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b2
7. Car prices prediction – EDA. Kaggle. Oct 26, 2019 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/code/kimyriel/car-prices-prediction-eda/notebook

Гіжевський Владислав Віталійович – студент групи 2ІСТ-22м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: vladgiz2000@gmail.com

Жуков Сергій Олександрович– к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Hizhevskyi Vladislav V. - student of Faculty of Intelligent Information Technology and Automation, 2IST-22m, Vinnytsia National Technical University, Vinnytsia, e-mail vladgiz2000@gmail.com

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com