

Ростислав Бачало, Світлана Яцишин*

¹ НЛТУ України, Львів, Україна, ORCID: 0009-0007-7776-1317,
bachalo.rostyslav@nltu.lviv.ua

² НЛТУ України, Львів, Україна, ORCID: 0000-0001-5200-4837,
yatsyshyn@nltu.edu.ua

Порівняння регресійних моделей у задачі прогнозування вартості автомобіля за допомогою ML.Net

Анотація. Дослідження присвячене прогнозуванню вартості автомобіля за допомогою машинного навчання на платформі ML.Net на основі їхніх технічних характеристик, таких як модель, рік випуску, тип палива, пробіг, об'єм двигуна та витрата палива. В дослідженні використано кілька алгоритмів регресії, таких як LightGbmRegression, FastTreeTweedieRegression, FastTreeRegression, SdcaRegression і LbfgsPoissonRegression. Моделі були оцінені за метриками R^2 (коефіцієнт детермінації), absolute-loss (абсолютна похибка), squared-loss (квадратична похибка) та RMS-loss (середньоквадратичне похибка). Найкращі результати отримані для моделей LightGbmRegression та FastTreeRegression, що підтверджує їхню високу точність у прогнозуванні вартості транспортних засобів. Дослідження відкриває нові можливості для подальшого вдосконалення моделей та інтеграції в реальні бізнес-процеси.

Ключові слова – прогнозування цін, машинне навчання, регресія, штучний інтелект, ML.Net.

Модель LightGbmRegression є однією з найбільш продуктивних у завданнях регресії. Вона базується на градієнтному бустингу над деревами рішень, що дає змогу досягти високої точності за рахунок ітеративного покращення моделі. Цей підхід забезпечує моделі здатність швидко адаптуватися до складних залежностей у даних [1]. У нашому випадку LightGbmRegression продемонструвала один із найкращих результатів із показником ($R^2 = 0.9456$) та низьким значенням середньоквадратичної похибки ($RMS = 1106.16$) та абсолютної похибки (796.99), що свідчить про її здатність точно передбачати вартість автомобілів на основі технічних характеристик.

Модель FastTreeTweedieRegression використовує алгоритм дерева рішень, орієнтований на розв'язання задач регресії за допомогою функції втрат Tweedie[1-2]. Ця модель є ефективною в умовах, коли залежності між змінними можуть бути складними та нелінійними. FastTreeTweedieRegression показала показник ($R^2 = 0.9406$) з невеликою абсолютною похибкою (810.83) і також відзначилася високою точністю та стабільністю прогнозів.

SdcaRegression (Stochastic Dual Coordinate Ascent) є моделлю, яка відома своєю ефективністю та швидкістю в задачах лінійної регресії. Вона працює шляхом оптимізації дуальних координат для знаходження мінімуму функції втрат.[1-2]. Хоча SdcaRegression продемонструвала нижчий показник точності, зокрема($R^2 = 0.7629$), її перевага полягає в швидкій обробці великих наборів даних, що може бути корисним у сценаріях, де критично важлива швидкість

LbfgsPoissonRegression (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) є методом оптимізації, який застосовується до задач логістичної та пуассонової регресії. Ця модель відзначається своєю здатністю працювати з великими наборами даних за рахунок ефективного використання пам'яті [1-2], але в нашому дослідженні вона показала дещо нижчі результати порівняно з іншими моделями ($R^2 = 0.8903$).

Найкращі результати отримані моделями LightGbmRegression і FastTreeRegression, які продемонстрували високі показники точності. Обидві моделі також показали низькі значення похибок (RMS 1106.16 для LightGbm та RMS 1119.53 для FastTree), що робить їх ефективними для прогнозування ринкової вартості автомобілів. Інші моделі, такі як SdcaRegression, показали гірші результати за точністю, проте відрізнялись більшою швидкістю оброблення даних, що може бути корисним у задачах, де швидкість є критичним параметром. Тестування моделей також виявило, що моделі з високим коефіцієнтом детермінації здатні забезпечити стабільні результати при роботі з великими наборами даних. Це дає змогу інтегрувати їх у реальні процеси автомобільної промисловості для покращення оцінки автомобілів при купівлі, продажу або страхуванні. Усі результати зображені на рис.1.

Add CarPricePrediction2.0ML.Model					
	Trainer	RSquared	Absolute-loss	Squared-loss	RMS-loss
1	SdcaRegression	0.7629	1772.96	5329386.09	2308.55
2	LightGbmRegression	0.9456	796.99	1223583.09	1106.16
3	FastTreeRegression	0.9442	796.62	1253346.60	1119.53
4	FastTreeTweedieRegression	0.9406	810.83	1335736.13	1155.74
5	FastForestRegression	0.8586	1321.55	3178151.88	1782.74
6	LbfgsPoissonRegression	0.8903	1156.92	2465560.64	1570.21
7	OnlineGradientDescentRegression	0.4694	2640.33	11927197.69	3453.58
8	OlsRegression	0.8614	1340.28	3115035.74	1764.95
9	LightGbmRegression	0.9442	803.35	1255334.93	1120.42
10	FastTreeRegression	0.9419	812.00	1305412.87	1142.55
11	FastTreeTweedieRegression	0.9438	798.53	1262436.95	1123.58
12	LightGbmRegression	0.9377	870.70	1400805.92	1183.56
13	FastTreeRegression	0.9067	1042.50	2097085.29	1448.13
14	FastTreeTweedieRegression	-4.3925	10180.70	121227183.79	11010.32
15	LightGbmRegression	0.9353	872.32	1454538.75	1206.04
16	FastTreeRegression	-2.3444	7988.31	75183787.30	8670.86
17	FastTreeTweedieRegression	-6.6543	12231.72	172072789.65	13117.65
18	LightGbmRegression	0.9024	1074.39	2193549.59	1481.06
19	FastTreeRegression	0.9453	780.32	1228613.25	1108.43
20	FastTreeTweedieRegression	-0.3669	4783.41	30729193.08	5543.39
21	LightGbmRegression	0.9317	893.70	1536269.38	1239.46

Рисунок 1. Результати тренувань моделей

Висновок. Дослідження продемонструвало, що моделі LightGbmRegression і FastTreeRegression є найбільш ефективними для прогнозування вартості автомобілів, забезпечуючи високу точність та надійність результатів. Їх можна використовувати в реальних сценаріях для оцінки транспортних засобів на ринку.

Список використаних літературних джерел

- [1] Ke, Guolin, et al. "LightGBM: A highly efficient gradient boosting decision tree." Advances in neural information processing systems, 2017.
- [2] Bishop, C. Pattern Recognition and Machine Learning. – Springer, 2006.
- [3] Документація ML.NET [Електронний ресурс] – режим доступу до ресурсу:» <https://dotnet.microsoft.com/en-us/apps/machinelearning-ai/ml-dotnet>«

Comparison of regression models in the task of predicting the cost of a car using ML.Net

Rostyslav Bachalo, Svitlana Yatsyshyn

The research is devoted to predicting the cost of a car using machine learning based on the ML.Net framework. Based on their technical characteristics, such as model, year of manufacture, fuel type, mileage, engine size, and fuel consumption. Several regression algorithms were used, such as LightGbmRegression, FastTreeTweedieRegression, FastTreeRegression SdcaRegression and LbfgsPoissonRegression. The models were evaluated by R^2 (coefficient of determination), absolute-loss,

squared-loss, and RMS-loss metrics. The best results were obtained for the LightGbmRegression and FastTreeRegression models, which confirms their high accuracy in predicting the value of vehicles. The study opens up new opportunities for further improving the models and integrating them into real business processes.