УДК 330.43 **DOI: 10.26565/2311-2379-2019-97-04**

Е.О. Ковпак, Ф.І. Орлов
Харківський національний університет імені В. Н. Каразіна
пл. Свободи, 4, м. Харків, 61022, Україна
E-mail: elvira.kovpak@karazin.ua, zzoorroo97@gmail.com
ORCID: https://orcid.org/0000-0001-9236-3084, https://orcid.org/0000-0002-3503-9832

ПОРІВНЯЛЬНИЙ АНАЛІЗ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ І РЕГРЕСІЙ ДЛЯ ПРОГНОЗУВАННЯ ЦІНИ ЛЕГКОВОГО АВТО

Метою дослідження, описаного у цій статті, є порівняльний аналіз прогнозних якостей деяких моделей машинного навчання та регресій, в яких факторами виступають споживчі характеристики вживаного легкового автомобіля: марка автомобіля, тип коробки передач, тип приводу, тип двигуна, пробіг, тип кузову, рік випуску, область продавця, стан авто, чи було авто у ДТП, середня ціна на аналог в Україні, об'єм двигуна, кількість дверей, наявність додаткового обладнання, кількість місць для пасажирів, чи перша реєстрація авто, чи пригнане авто із закордону. Якісні змінні були закодовані як бінарні змінні або за допомогою середнього значення цільової змінної. Для моделювання було використано понад 200 тисяч автомобілів. Оцінка параметрів усіх моделей проводилася у середовищі Python із використанням бібліотек Sklearn, Catboost, StatModels та Keras. У ході дослідження були розглянуті такі моделі регресій та моделі машинного навчання: лінійна регресія; поліноміальна регресія; дерево рішень; нейронна мережа; моделі за алгоритмами «k-найближчих сусідів», «випадковий ліс», «градієнтний бустинг»; ансамбль моделей. У статті представлені найкращі з точки зору якості (згідно критеріїв R^2, MAE, MAD, MAPE) варіанти із кожного класу моделей. Було виявлено, що найкраще із задачею прогнозування ціни на легковий автомобіль справляються саме нелінійні моделі. Результати моделювання свідчать про те, що найкраще відображає залежність між ціною легкового автомобіля та його характеристиками саме ансамбль моделей, до якого увійшли нейронна мережа, моделі за алгоритмами «випадковий ліс» та «градієнтний бустинг». Ансамбль моделей показав середню відносну похибку апроксимації вихідних даних 11,2%, та середню відносну похибку прогнозу 14,34%. Усі запропоновані нелінійні моделі ціни на авто мають приблизно однакові прогнозні якості (різниця між MAPE у межах 2%).

Ключові слова: ціна автомобіля, регресія, нейронні мережі, ансамбль моделей.
JEL Classification: C45, C51, C52, C55.

Elvira Kovpak, Fedir Orlov
V.N. Karazin Kharkiv National University
4 Svobody Sq., 61022, Kharkiv, Ukraine
E-mail: elvira.kovpak@karazin.ua, zzoorroo97@gmail.com
ORCID: https://orcid.org/0000-0001-9236-3084, https://orcid.org/0000-0002-3503-9832

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS AND REGRESSIONS FOR CAR PRICE PREDICTION

The purpose of the research described in this article is a comparative analysis of the predictive qualities of some models of machine learning and regression. The factors for models are the consumer characteristics of a used car: brand, transmission type, drive type, engine type, mileage, body type, year of manufacture, seller's region in Ukraine, condition of the car, information about accident, average price for analogue in Ukraine, engine volume, quantity of doors, availability of extra equipment, quantity of passenger's seats, the first registration of a car, car was driven from abroad or not. Qualitative variables has been encoded as binary variables or by mean target encoding. The information about more than 200 thousand cars have been used for modeling. All models have been evaluated in the Python Software using Sklearn, Catboost, StatModels and Keras libraries. The following regression models and machine learning models were considered in the course of the study: linear regression; polynomial regression; decision tree; neural network; models based on "k-nearest neighbors", "random forest", "gradient boosting" algorithms; ensemble of models. The article presents the best in terms of quality (according to the criteria R^2, MAE, MAD, MAPE) options from each class of models. It has been found that the best way to predict the price of a passenger car is through non-linear models. The results of the modeling show that the dependence between the price of a car and its characteristics is best described by the ensemble of models, which includes a neural network, models using "random forest" and "gradient boosting" algorithms. The

ensemble of models showed an average relative approximation error of 11.2% and an average relative forecast error of 14.34%. All nonlinear models for car price have approximately the same predictive qualities (the difference between the MAPE within 2%) in this research.

Keywords: car price, regression, neural networks, ensemble of models.
JEL Classification: C45, C51, C52, C55.

Э.А. Ковпак, Ф.И. Орлов
Харьковский национальный университет имени В. Н. Каразина
пл. Свободы, 4, г. Харьков, 61022, Украина
E-mail: elvira.kovpak@karazin.ua, zzoorroo97@gmail.com
ORCID: https://orcid.org/0000-0001-9236-3084, https://orcid.org/0000-0002-3503-9832

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ И РЕГРЕССИЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ЦЕНЫ ЛЕГКОВОГО АВТО

Целью исследования, описанного в этой статье, является сравнительный анализ прогнозных качеств некоторых моделей машинного обучения и регрессий, в которых факторами выступают потребительские характеристики подержанного легкового автомобиля: марка автомобиля, тип коробки передач, тип привода, тип двигателя, пробег, тип кузова, год выпуска, область продавца, состояние авто, было авто в ДТП, средняя цена на аналог в Украине, объем двигателя, количество дверей, наличие дополнительного оборудования, количество мест для пассажиров, сведения о регистрации авто, пригнан ли автомобиль из-за рубежа. Качественные переменные были закодированы как бинарные переменные или с помощью среднего значения целевой переменной. Для моделирования были использованы более 200 тысяч автомобилей. Оценка параметров всех моделей проводилась в среде Python с использованием библиотек Sklearn, Catboost, StatModels и Keras. В исследовании были рассмотрены такие модели регрессий и модели машинного обучения: линейная регрессия; полиномиальная регрессия; дерево решений; нейронная сеть; модели с алгоритмами «k-ближайших соседей», «случайный лес», «градиентный бустинг»; ансамбль моделей. В статье представлены лучшие с точки зрения качества (согласно критериев R^2, MAE, MAD, MAPE) варианты из каждого класса моделей. Было выявлено, что лучше всего с задачей прогнозирования цены на легковой автомобиль справляются именно нелинейные модели. Результаты моделирования свидетельствуют о том, что лучше всего отражает зависимость между ценой легкового автомобиля и его характеристиками именно ансамбль моделей, в который вошли нейронная сеть, модели по алгоритмам «случайный лес» и «градиентный бустинг». Ансамбль моделей показал среднюю относительную погрешность аппроксимации исходных данных 11,2%, и среднюю относительную погрешность прогноза 14,34%. Все предложенные нелинейные модели цены на авто имеют примерно одинаковые прогнозные качества (разница между MAPE в пределах 2%).

Ключевые слова: цена автомобиля, регрессия, нейронные сети, ансамбль моделей.
JEL Classification: C45, C51, C52, C55.

Постановка проблеми. Оцінка ринкової вартості автомобіля передує будь-якій угоді на автомобільному ринку. Зазвичай оцінка ринкової вартості вживаного автомобіля спирається на експертні оцінки або метод аналогій. Використання формалізованої процедури оцінки ринкової вартості вживаного автомобіля за умов її автоматизації надасть змогу учасникам ринку легко визначати очікувану ціну на авто для операцій купівлі-продажу, кредитування під заставу, страхування, вирішення майнових спорів та ін. Розробка моделі визначення ринкової ціни на автомобіль є проблемою прикладного характеру.

Машинне навчання – галузь знань, що динамічно розвивається і затребувана завдяки можливості інтеграції моделей машинного навчання в експертні системи та інформаційні ресурси за умов переоцінки або налаштування параметрів на підставі регулярної процедури оновлення навчальної вибірки. Цікавим представляється порівняння можливостей економетричних методів та методів машинного навчання для прогнозування ціни вживаного легкового авто.

Аналіз останніх досліджень і публікацій. Серед формалізованих методів прогнозування цін найбільш популярними методами є економетричні методи прогнозування ціни (Осокина, 2015), (Журкина, 2015), (Мрочко & Батожаргалов, 2015), (Валеева & Исавнин, 2016), (Утакаева, 2019). В економетричних моделях найчастіше використовують лише декілька основних факторів для визначення ринкової ціни автомобіля: пробіг, рік випуску, об'єм двигуна, тип палива, тип кузова та тип коробки передач автомобіля (Журкина, 2015), (Мрочко &

Батожаргалов, 2015), (Валеева & Исавнин, 2016). У дослідженні Валеєва З.Ф та Ісавніна А.Г (Валеева & Исавнин, 2016) для моделювання ціни автомобіля використовується 20 факторів.

Моделі машинного навчання для визначення цін на автомобілі запропоновано в роботах (Gegic, Isakovic, Keco, Masetic & Kevric, 2019), (Kanwal & Sadaqat, 2017), (Ozcalici, 2017). Зокрема, у дослідженні (Gegic et al., 2019) використовуються моделі, побудовані за алгоритмами «випадковий ліс» та «модель опірних векторів» (SVM), нейронну мережу та ансамбль моделей.

Метою дослідження є побудова та порівняльний аналіз регресійних моделей та моделей машинного навчання, що використовують споживчі характеристики вживаних легкових автомобілів для оцінки їх ринкової вартості.

База дослідження. Для побудови моделей використовувались емпіричні дані – оголошення про продаж автомобілів з ресурсу autoria.ua (Autoria, 2019) (вибірка із 200 939 об'єктів). Для побудови та оцінки моделей було використано мову програмування Python та бібліотеки для реалізації методів машинного навчання StatModels, Sklearn, Catboost та Keras (Statmodels), (Sklearn), (Catboost), (Keras). У навчальну вибірку увійшло 90% від загальної вибірки автомобілів із оголошень, що досліджувались (180845 об'єктів), а у тестову увійшло 10% авто (20094 об'єктів).

Основні результати дослідження. У ході роботи було побудовано декілька варіантів кожної розглянутої моделі машинного навчання з різними факторами, параметрами та архітектурою. У цій статі представлені лише найкращі з огляду на якість на тестовій вибірці варіанти кожної моделі.

Спочатку було здійснено спробу отримати адекватну лінійну регресійну модель на підставі незалежних ціноутворюючих факторів. Фактори, що в дослідженні визначають ринкову вартість автомобілю, було поділено на числові та категоріальні (латинське позначення перших написано маленькими, а других – великими літерами) (табл. 2).

Таблиця 2

Перелік ціноутворюючих факторів

Споживча характеристика авто	Тип фактору	Назва в моделі	Область можливих значень	Приклад
Марка автомобіля	категоріальний	PRODUCER	[0; ∞]	Ford
Ручна коробка передач	категоріальний	TRANSMISSION	{0, 1}	Автомат
Привід	категоріальний	POWERTRAIN	{Передний, Полный}	Передній
Пробіг	числовий	mileage	[0; ∞]	100000
Тип двигуна	категоріальний	FUEL	{дизель, електро, газ, газ-бензин, газ-метан, газ-пропан-бутан, гібрид, інше}	Дизельний
Рік випуску	числовий	year	[1930; 2019]	2008
Область продавця	категоріальний	REGION	[0; ∞]	Київська
Тип кузова	категоріальний	BODY	{Кабриолет, Купе, Легковой фургон, Лимузин, Лифтбек, Минивэн, Пикап, Родстер, Седан, Универсал, Хэтчбек }	Седан
Чи був у ДТП	категоріальний	DPT	{0, 1}	Ні
Додаткове обладнання	числовий	condition_n, safety_n, comfort_n, multimedia_n, other_n, total_features	[0; ∞]	Обігрів сидінь, GPS, парктронік
Середня ціна	числовий	model_year	[0; ∞]	9000
Стан авто	категоріальний	MY_ESTIMATION	{1, 2, 3}	Добрий
Об'єм двигуна	числовий	volume	[0; ∞]	3.5
Кількість дверей	числовий	doors	[0; ∞]	4
Кількість місць для пасажирів	числовий	seats	[0; ∞]	5
Перша реєстрація	категоріальний	FIRST_REG	{0, 1}	Так
Пригнаний із-за кордону	категоріальний	PRIGNANA	{0, 1}	Ні

Джерело: авторська розробка

Представлений перелік факторів не є вичерпним, тому що надалі у окремих моделях будуть вводитися нові похідні фактори та їх позначення, які будуть створюватися за допомогою початкових, що перелічені у табл. 2. Значимість оцінок параметрів біля факторів перевірялася за допомогою t-критерію на підставі значення рівня значущості (p-значенню). Рівень значущості оцінок параметрів біля факторів в моделях не нижчий за 5%. Деякі категоріальні фактори (а саме PRODUCER та REGION) були закодовані за допомогою середнього значення цільової змінної (mean target encoding) (Castillo, 2019): кожне якісне значення категоріального фактору було закодоване відповідним середнім числовим для нього по навчальній вибірці.

Для порівняння різних за своєю будовою та структурую моделей було обрано чотири критерії якості моделювання, які можуть бути застосовані до усіх моделей, що розглядаються у цій роботі, а саме: коефіцієнт детермінації (R^2), середня абсолютна похибка (MAE), медіанна абсолютна похибка (MAD), середня абсолютна відсоткова помилка (MAPE).

Для реалізації мети дослідження було побудовано такі моделі: лінійна регресія; поліноміальна регресія; дерево рішень; нейронна мережа; моделі за алгоритмами «k-найближчих сусідів», «випадковий ліс», «градієнтний бустинг»; ансамбль моделей.

Більшість із вказаних моделей мають параметри, які задаються до початку побудови моделі. Ці параметри визначаються складність моделі та деякі спеціальні характеристики, що вказуються у таблицях у стовпчику «Параметри».

Рівняння множинної лінійної регресії із оціненими коефіцієнтами має такий вид:

$$y = -1.106*10^5 + 0.1867*PRODUCER + 53.3793*year - 7.6064*mileage + 1198.7728*volume + 1055.2298*TRANSMISSION + 0.0538*REGION - 211.6692*seats + 0.9407*model\_year - 335.5988*PRIGNANA - 4273.3431*DPT + 252.3447*POWERTRAIN\_Передний + 148.7327*POWERTRAIN\_Полный + 5165.1986*BODY\_Кабриолет + 3060.7428*BODY\_Купе + 565.6259*BODY\_Легковой фургон (до 1,5 т) + 3.993*10^4*BODY\_Лимузин + 166.9068*BODY\_Лифтбек + 1366.9044*BODY\_Минивэн + 747.6891*BODY\_Пикап + 3943.5985*BODY\_Родстер + 898.7058*BODY\_Седан + 1121.1102*BODY\_Универсал + 1073.1227*BODY\_Хэтчбек \quad (1)$$

Модель лінійної регресії (1) показала незадовільну прогнозну якість на тестовій вибірці (див. табл. 3). Це може бути пов'язано із принципіальною неспроможністю лінійної моделі враховувати нелінійні зв'язки між залежною змінною та незалежними.

Таблиця 3

Показники точності множинної лінійної регресії (1)

Число параметрів	Найменування факторів	Вибірка	R^2	MAE	MAD	MAPE
24	PRODUCER, year, mileage, volume, REGION, seats, model_year, PRIGNANA, DPT, TRANSMISSION, POWERTRAIN, BODY	Навчальна	0.75	2274	1124	35.18
		Тестова	0.64	2530	1172	39.39

Джерело: авторська розробка

Для поліноміальної регресії ціни на вживане авто були використані деякі фактори із табл.2 у другому степені та їх інтеракції (змінні, що представляють собою комбінації інших факторів – сума, різниця, добуток або частка двох або більше регресорів). У результаті відбору факторів та оцінки параметрів поліноміальної регресії було отримано таке рівняння:

$$y = 2.697*10^7 + 0.1854*PRODUCER - 2.691*10^4*year + 14.8008*mileage - 2.402*10^5*volume + 253.0908*TRANSMISSION + 0.0491*REGION + 6531.2280*total\_features - 24.5455*model\_year - 352.2857*PRIGNANA - 4649.5836*DPT - 359.3765*POWERTRAIN\_Передний + 441.4734*POWERTRAIN\_Полный + 5015.3232*BODY\_Кабриолет + 2805.4425*BODY\_Купе + 787.2248*BODY\_Легковой фургон (до 1,5 т) + 3.935e+04*BODY\_Лимузин - 346.4346*BODY\_Лифтбек + 1256.6041*BODY\_Минивэн - 932.1227*BODY\_Пикап + 4752.2303*BODY\_Родстер + 252.3842*BODY\_Седан + 658.6929*BODY\_Универсал + 539.7720*BODY\_Хэтчбек + 6.7130*year^2 + 0.0075*mileage^2 + 237.5860*volume^2 + 3.3530*total\_features^2 + 4.661*10^{-8}*model\_year^2 + 120.8433*year\_volume - 3.2819*year\_total\_features + 0.0125*year\_model\_year - 8.7004*mileage\_volume + 0.1751*mileage\_total\_features - 0.0009*mileage\_model\_year - 38.7573*volume\_total\_features + 0.0216*volume\_model\_year + 0.0024*total\_features\_model\_year \quad (2)$$

Із таблиці 4 видно, що додавання поліноміальної частини та інтеракцій факторів, тобто спроба урахувати нелінійність зв'язків між залежною та незалежними змінними, дало значний приріст у якості на тестовій множині. Так, наприклад, середня похибка за показником MAPE впала більше ніж на 11%.

Таблиця 4

Показники якості поліноміальної регресії (2)

Число параметрів	Фактори	Вибірка	R^2	MAE	MAD	MAPE
38	PRODUCER, year, mileage, volume, TRANSMISSION, REGION, total_features, model_year, PRIGNANA, DPT, POWERTRAIN, BODY, year^2, mileage^2, volume^2, total_features^2, model_year^2, year_volume, year_total_features, year_model_year, mileage_volume, mileage_total_features, mileage_model_year, volume_total_features, volume_model_year, total_features_model_year	Навчальна	0.76	2155	990	26.76
		Тестова	0.65	2285	1010	27.95

Джерело: авторська розробка

Розглянемо модель машинного навчання за алгоритмом «k-найближчих сусідів» (KNN). Усі фактори були стандартизовані за формулою:

$$x' = \frac{x - \bar{x}}{\sigma},$$

де x' – нове значення фактору

x – старе значення фактору

$\bar{x}$ – середнє значення фактора

σ – стандартне відхилення фактору.

Розрахунок середнього значення фактору та його стандартного відхилення реалізується на навчальній вибірці. Таке перетворення факторів дозволяє компенсувати різний масштаб змінних та, відповідно, різний вплив на величину відстані.

Задати значимість факторів можна, якщо просто збільшити усі значення фактору на якесь число. Коефіцієнти, на які домножуються значення факторів, були підібрані так, щоб відповідати ранжуванню значимості факторів у лінійних моделях та теоретичним передумовам (табл. 5).

Таблиця 5

Зміна значень деяких факторів моделі KNN

Фактор	Коефіцієнт
PRODUCER	3
year	1.5
mileage	2
volume	2
model_year	7

Джерело: авторська розробка

Показники якості побудованої за алгоритмом «k-найближчих сусідів» моделі представлено у табл. 6.

Таблиця 6

Показники точності моделі KNN

Число факторів / вимірів	Фактори	Параметри	Вибірка	R^2	MAE	MAD	MAPE
19	PRODUCER, year, mileage, volume, TRANSMISSION, REGION, total_features, model_year, DPT, POWERTRAIN, year_volume, year_model_year, mileage_PRODUCER, volume_model_year, volume_PRODUCER	n_neighbors = 22, weights=' distance'	Навчальна	1.0	34	0	0.04
			Тестова	0.79	1674	644	16.3

Джерело: авторська розробка

Із табл. 6 можна зробити висновок, що модель за алгоритмом «k-найближчих сусідів» набагато краще моделює ціну на авто ніж лінійна та поліноміальна регресія. Так, середня похибка за показником MAPE знову впала більше ніж на 11%.

Наступною моделлю, що розглядалася для моделювання ціни вживаного авто, є дерево рішень. Для моделі дерева рішень були спробувані різні способи обмеження складності, що дозволило побудувати найкращий варіант із якістю та параметрами, що представлено у табл. 7.

Таблиця 7

Показники точності моделі дерева рішень

Число параметрів	Найменування факторів	Параметри	Вибірка	R^2	MAE	MAD	MAPE
43	PRODUCER, year, mileage, volume, REGION, doors, seats, total_features, model_year, FIRST_REG, PRIGNANA, DPT, condition_n, safety_n, comfort_n, multimedia_n, other_n, MY_ESTIMATION, POWERTRAIN, FUEL, BODY, TRANSMISSION	max_depth = 30, min_samples _split = 40, min_samples _leaf = 20	Навчальна	0.79	1475	546	13.1
			Тестова	0.85	1786	657	16.3

Джерело: авторська розробка

Порівнюючи таблиці 6 та 7 можна прийти до висновку, що дерево рішень показує себе краще за модель за алгоритмом «k-найближчих сусідів» лише за одним критерієм – R^2, бо дозволяє краще прогнозувати ринкову ціну на автомобілі з нестандартними споживчими характеристиками.

Далі була побудована модель ціни на авто згідно алгоритму «випадкового лісу», яка у свою чергу складається із 150 окремих дерев рішень. Таке комбінування дерев рішень

дозволило значно покращити якість прогнозування відносно попередніх моделей ціни на авто. Параметри моделі, її фактори та показники якості представлено у табл. 8.

Таблиця 8

Показники якості моделі ціни на авто за алгоритмом «випадковий ліс»

Число параметрів	Найменування факторів	Параметри	Вибірка	R^2	MAE	MAD	MAPE
43	PRODUCER, year, mileage, volume, REGION, doors, seats, total_features, model_year, FIRST_REG, PRIGNANA, DPT, condition_n, safety_n, comfort_n, multimedia_n, other_n, MY_ESTIMATION, POWERTRAIN, FUEL, BODY, TRANSMISSION	n_estimators =150, max_features =0.85, random_state=0, max_depth = 34, min_samples_split = 10, min_samples_leaf = 5	Навчальна	0.86	1031	378	9.43
			Тестова	0.84	1552	586	14.5

Джерело: авторська розробка

Наступною розглядалася модель за алгоритмом «градієнтного бустингу». Показники якості, параметри та фактори найкращого варіанту моделі знаходяться у таблиці 9.

Таблиця 9

Показники точності моделі за алгоритмом «градієнтного бустингу»

Кіл-сть параметрів	Фактори	Параметри	Вибірка	R^2	MAE	MAD	MAPE
33	PRODUCER, year, mileage, volume, FUEL, TRANSMISSION, POWERTRAIN, BODY, REGION, doors, seats, total_features, model_year, FIRST_REG, PRIGNANA, DPT, condition_n, safety_n, comfort_n, multimedia_n, other_n, MY_ESTIMATION, year_mileage, year_volume, year_model_year, mileage_volume, mileage_model_year, volume_model_year, year^2, mileage^2, volume^2, total_features^2, model_year^2	learning_rate=0.05, l2_leaf_reg=0, bagging_temperature = 3, border_count=1000, depth=13, iterations=1100, random_seed=0, eval_metric=MAE, , random_strength =4, rms=0.7	Навчальна	0.99	881	539	12.18
			Тестова	0.91	1521	624	15.41

Джерело: авторська розробка

Далі була побудована нейронна мережа. Найважливішими параметрами для нейронної мережі є кількість шарів, нейронів у шарах, використані функції активації, регуляризація та алгоритм оптимізації. Архітектура найкращого варіанту нейронної мережі зображено на рис.1,

показники якості, параметри та фактори, що були використані для моделювання ціни, зображені у табл. 10.

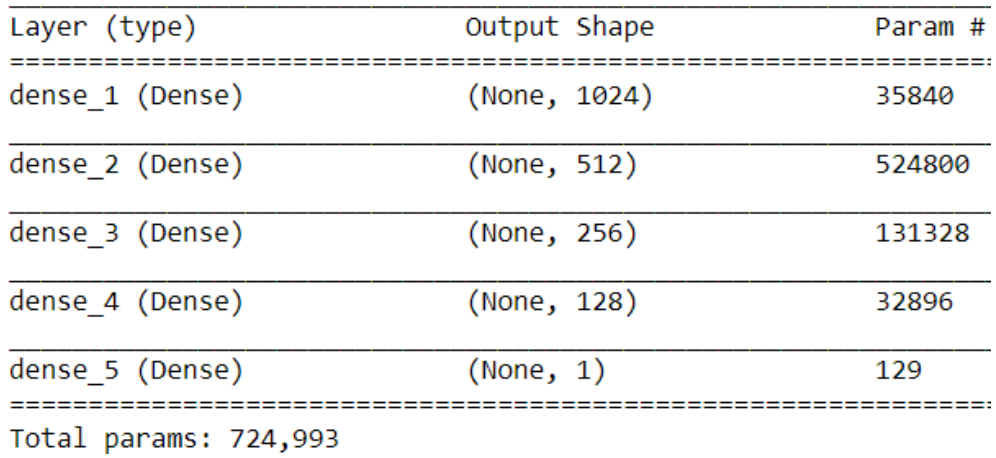

Рис. 1. Архітектура нейронної мережі

Джерело: авторська розробка

Таблиця 10

Показники точності нейронної мережі

Кіл-сть входів	Фактори	Параметри	Вибірка	R^2	MAE	MAD	MAPE
34	PRODUCER, year, mileage, volume, TRANSMISSION, REGION, seats, total_features, model_year, PRIGNANA, DPT, POWERTRAIN, BODY, year_mileage, year_volume, year_model_year, mileage_volume, mileage_model_year, volume_model_year, year^2, mileage^2, volume^2, model_year^2	optimizer = RMSprop, learning_rate=0.03, activation = RELu, linear (ост.шар), batch_size=128, epochs =10, регуляризація = L1(0.01), L2(0.01)	Навчальна	0.84	1472	603	13.94
			Тестова	0.82	1654	635	15.35

Джерело: авторська розробка

Із аналізу табл. 10 можна зробити висновок, що критерії якості апроксимації та прогнозування нейронної мережі наближені до аналогічних показників у моделі за алгоритмом «випадковий ліс».

Останньою моделлю ринкової ціни на авто розглядався ансамбль моделей – поєднання окремих моделей для отримання кращого результату. Техніка комбінації кількох моделей застосовується і таких моделях, як моделі за алгоритмами «градієнтний бустинг» та «випадковий ліс». Найпростіший спосіб поєднання кількох моделей в одну для отримання комбінованого прогнозу – це розрахунок простого середнього прогнозів серед прогнозів за обраними моделями. Саме таким чином і був створений ансамбль моделей, що включає у себе нейронну мережу, моделі за алгоритмами «випадковий ліс» та «градієнтний бустинг». Результати якості ансамблю моделей показані у таблиці 11.

Таблиця 11

Показники якості для ансамблю моделей

Моделі, що увійшли до ансамлю	Вибірка	R^2	MAE	MAD	MAPE
«Випадковий ліс», «Градієнтний бустинг», Нейронна мережа	Навчальна	0.93	1067	486	11.2
	Тестова	0.88	1492	587	14.34

Джерело: авторська розробка

Отже, в дослідженні було отримано 7 альтернативних моделей ціни на вживане легкове авто. Результати порівняльного аналізу прогнозних якостей моделей представлено у табл. 12.

Таблиця 12

Порівняння прогнозних якостей моделей (на тестових вибірках)

Модель	R^2	MAE	MAD	MAPE
Лінійна регресія	0.64	2530	1172	39.39
Поліноміальна регресія	0.65	2285	1010	27.95
KNN	0.79	1674	644	16.25
Дерево рішень	0.85	1786	657	16.25
«Випадковий ліс»	0.84	1552	586	14.5
«Градієнтний бустинг»	0.91	1521	624	15.41
Нейронна мережа	0.82	1654	635	15.35
Ансамбль	0.88	1492	587	14.34

Джерело: авторська розробка

Таким чином, модель ансамблю є найкращою моделлю по усіх критеріях якості апроксимації, окрім R^2. Показник MAD на 1 одиницю виявився гіршим ніж для ансамблю, ніж у моделі за алгоритмом «випадковий ліс», але можна вважати таку різницю незначимою. Деяке нехтування показником R^2 пов'язане з тим, що цей критерій сильно реагує на екстремальні помилки для нетипових об'єктів (викидів). А так як акцент побудови моделі ринкової вартості ціни на авто ставиться саме на її узагальнюючій можливості, то набагато важливішим є краще пристосування моделі до типових та найпоширеніших представників вибірки. Якісна перевага ансамбля моделей пояснюється тим, що в ансамбль було включено три найкращі із семи побудованих моделей. Отже, саме машинні методи навчання показали найкращі прогнозні результати якості для визначення ціни на легковий вживаний автомобіль.

Висновки. Результатами проведеної роботи можна вважати наступе:

1) виявлено, що нелінійні моделі значно краще справляються із задачею моделювання ціни на легковий автомобіль, ніж лінійні моделі. Це може пояснюватися наявністю нелінійного зв'язку між ціною автомобіля та його споживчими характеристиками;

2) усі розглянуті нелінійні моделі мають приблизно однакові прогнозні якості (різниця між середньою відносною похибкою апроксимації найбільш і найменш точних нелінійних моделей у межах 2%);

3) саме машинні методи навчання показали найкращі прогнозні якості для визначення ціни на легковий вживаний автомобіль порівняно із регресіями;

4) комбінація кількох методів моделювання ціни авто в рамках ансамблю моделей (алгоритми «випадковий ліс», «градієнтий бустинг» та нейронна мережа) показала найкращі якість апроксимації вихідних даних та прогнозну здатність;

5) реалізація прогнозування ціни на легковий автомобіль на підставі отриманого ансамблю моделей може бути автоматизована для створення експертної системи для оцінки очікуваної ринкової вартості вживаного авто, що дає змогу за лічені хвилини отримувати формалізований прогноз ціни.

Література

1. Осокина О.А. Эконометрическое моделирование стоимости автомобиля Renault Duster на вторичном рынке. *VII Международная студенческая научная конференция «Студенческий научный форум»*. Москва, 2015. С. 8.
2. Журкина Е.А. Эконометрическое моделирование стоимости автомобиля Renault Duster на вторичном рынке. *VII Международная студенческая научная конференция «Студенческий научный форум»*. Москва, 2015. С. 6.
3. Мрочко А.А., Батожаргалов Б. Эконометрический анализ рынка подержанных автомобилей (BMW, MERCEDES, AUDI). *Международный студенческий научный вестник*. Пенза, 2015. №6. С. 28.
4. Валеева З.Ф., Исавнин А.Г. Эконометрическое моделирование стоимости автомобилей на вторичном рынке в городе Набережные Челны. *Фундаментальные исследования*. Москва, 2016. № 6. С. 154–158.
5. Утакаева И. Х. Применение пакета статистического анализа Python для анализа данных автомобильного рынка. *Вестник Алтайской академии экономики и права*. 2019. № 2. С. 346–351.
6. Gegic E., Isakovic B., Keco D., Masetic Z. & Kevric J. Car Price Prediction using Machine Learning Techniques. Сараєво*: TEM Journa,* 2019. №1 (8). С. 113–118.
7. Kanwal N., Sadaqat J. Vehicle Price Prediction System using Machine Learning Techniques. *International Journal of Computer Applications*. Нью-Йорк, 2017. №9. Вип. 167. С. 27–31.
8. Ozcalici, M. Predicting Second-Hand Car Sales Price Using Decision Trees and Genetic Algorithms. *Alphanumeric Journal*. Стамбул, 2017. № 5. С. 103–114.
9. Autoria. URL: https://auto.ria.com.
10. Statmodels. URL: https://www.statsmodels.org/.
11. Sklearn. URL: https://scikit-learn.org/.
12. Catboost. URL: https://catboost.ai/.
13. Keras. URL: https://keras.io/.
14. How to Build, Develop and Deploy a Machine Learning Model to predict cars price using Neural Networks. *Medium.* URL: https://medium.com/thelaunchpad/how-to-build-develop-and-deploy-a-machine-learning-model-to-predict-cars-price-using-neural-7f7439a37300.

Reference

1. Osokina, O. (2015). Econometric modeling of the cost of the car Renault Duster in the secondary market. Moscow: *VII International Student Scientific Conference "Student Scientific Forum". p.8.* (in Russian)
2. Zhurkina, E. (2015). Econometric modeling of the cost of the car Ford Fiesta in the secondary market, the example calculations. Moscow*: VII International Student Scientific Conference "Student Scientific Forum". p.6.* (in Russian)
3. Mrochko, A., Batojargalov, B. (2015). Econometric analysis of the market for used cars (BMW, Mercedes, Audi). Moscow: *International student scientific newsletter*, 6, 28. (in Russian)
4. Valeeva, Z., Isavnin, A. (2016). Econometric modeling of price car in the secondary market in the city of Naberezhnye Chelny. Moscow: *Fundamental research*, 6, 154-158. (in Russian)
5. Utakaeva, I.H. (2019). Experience of econometric modeling using the statistical analysis package Python*. Bulletin of the Altai Academy of Economics and Law,* 2, 346-351*. (*in Russian)
6. Gegic, E., Isakovic, B., Keco, D., Masetic, Z. & Kevric, J. (2019). Car Price Prediction using Machine Learning Techniques. Saraevo: *TEM Journal, 1 (8)*, 113-118.
7. Kanwal, N., Sadaqat, J. (2017). Vehicle Price Prediction System using Machine Learning Techniques. New York: *International Journal of Computer Applications*, 167, 27-31.
8. Ozcalici, M. (2017). Predicting Second-Hand Car Sales Price Using Decision Trees and Genetic Algorithms. Istanbul: *Alphanumeric Journal*, №5, 103-114.
9. Autoria. Retrieved from https://auto.ria.com.
10. Statmodels. Retrieved from https://www.statsmodels.org/.
11. Sklearn. Retrieved from https://scikit-learn.org/.
12. Catboost. Retrieved from https://catboost.ai/.
13. Keras. Retrieved from https://keras.io/.
14. How to Build, Develop and Deploy a Machine Learning Model to predict cars price using Neural Networks (2019)*. Medium.* Retrieved from https://medium.com/thelaunchpad/how-to-build-develop-and-deploy-a-machine-learning-model-to-predict-cars-price-using-neural-7f7439a37300.