УДК 004.043
DOI https://doi.org/10.32782/IT/2023-3-1

Наталія БОЙКО
кандидат економічних наук, доцент, доцент кафедри систем штучного інтелекту, Національний університет «Львівська політехніка», вул. Степана Бандери, 12, м. Львів, Україна, 79036, Nataliya.i.boyko@lpnu.ua
***ORCID:** 0000-0002-6962-9363*
***Scopus Author ID:** 57191967462*

Богдан ЛЕВИЦЬКИЙ
студент кафедри систем штучного інтелекту, Національний університет «Львівська політехніка», вул. Степана Бандери, 12, м. Львів, Україна, 79036, bohdan.levytskyi.knm.2018@lpnu.ua
***ORCID:** 0000-0002-0060-2381*

Бібліографічний опис статті: Бойко, Н., Левицький, Б. (2023). Алгоритми тренування та оцінки моделей машинного навчання для структурованого набору даних. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 3, 3–12, doi: https://doi.org/10.32782/IT/2023-3-1

АЛГОРИТМИ ТРЕНУВАННЯ ТА ОЦІНКИ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ СТРУКТУРОВАНОГО НАБОРУ ДАНИХ

В статті розглядається послідовний процес попереднього аналізу та обробки структурованих даних про будівельні транспортні засоби різних типів. Наведений алгоритм побудови моделей машинного навчання, зокрема таких як лінійна регресія, дерево прийняття рішень та випадковий ліс, оцінка якості отриманих моделей та продукуючих результатів. Робота описує дослідження сфери покупки та продажу авто на вторинному ринку з використанням сучасних технологій data mining. Основна мета цього дослідження – передбачити вартість транспортного засобу з використанням атрибутів, що сильно корелюють з ціною. Пропонується розглянути концепти ціноутворення побудувавши наступні моделі машинного навчання: з урахуванням ознак специфічних певних марок автомобілів, з урахування ознак специфічних для певних типів автомобілів, а також загальну модель, яка включає усі наявні в наборі ознаки. Моделі було побудовано на основі методів лінійної регресії та дерева рішень. Метою відбору алгоритмів машинного навчання була мінімізація похибок при прогнозуванні вартості, швидкість роботи, легкість інтерпретації отриманих результатів: на основі яких даних приймалося рішення та які дані найбільше впливають на формування вартості. Для мінімізації похибки прогнозування було проведено детальний аналіз даних та їх підготовку для кожного типу будівельного транспортного засобу. Проведено багато експериментів з різними методами для пошуку та видалення аномальних спостережень, для пошуку та використання найбільш важливих ознак, при цьому використовувалися такі методи, як Z-index, міжквартильний розмах, рекурсивне видалення ознак, пошук ознак на основі виявлення залежностей з використанням статистичних методів. Було проведено порівняльний аналіз результатів кожної з моделей, проаналізовано можливі причини тих чи інших результатів. Наведені проблеми, які виникають при вирішенні даної задачі регресійного типу – відбір даних, що якнайкраще узагальнюють систему формування вартості технічного транспортного засобу.

***Ключові слова:** машинне навчання, дані, алгоритм, обробка даних, регресійні моделі машинного навчання, лінійна регресія, дерево прийняття рішень, випадковий ліс.*

Nataliya BOYKO
Candidate of Economical Sciences, Associate Professor, Associated Professor at the Department of Artificial Intelligence, Lviv Polytechnic National University, 12 Stepana Bandery str., Lviv, Ukraine, 79036, Nataliya.i.boyko@lpnu.ua
***ORCID:** 0000-0002-6962-9363*
***Scopus Author ID:** 57191967462*

Bohdan LEVYTSKYI
Student at the Department of Artificial Intelligence, Lviv Polytechnic National University, 12 Stepana Bandery str., Lviv, Ukraine, 79036, bohdan.levytskyi.knm.2018@lpnu.ua
***ORCID:** 0000-0002-0060-2381*

To cite this article: Boyko, N., Levytskyi, B. (2023). Algorytmy trenuvania ta otsinky modelei mashynnoho navchania dlia strykturovanoho naboru danykh [Algorithms of learning and evaluation of machine learning models of structured data sets]. *Information Technology: Computer Science, Software Engineering and Cyber Security,* 3, 3–12, doi: https://doi.org/10.32782/IT/2023-3-1

ALGORITHMS OF LEARNING AND EVALUATION OF MACHINE LEARNING MODELS OF STRUCTURED DATA SETS

The article deals with the sequential process of preliminary analysis and processing structured data about various types of construction vehicles. The algorithm for building machine learning models, in particular such as linear regression, decision tree and random forest, assessment of the quality of the obtained models and producing results is presented. The work describes research in the field of buying and selling cars on the secondary market using modern data mining technologies. The main objective of this study is to predict vehicle value using attributes highly correlated with price. It is proposed to consider the concepts of pricing by building the following machine learning models: taking into account the characteristics of specific brands of cars, taking into account the characteristics specific to certain types of vehicles, as well as a general model that includes all the characteristics available in the set. Models were built based on linear regression and decision tree methods. The purpose of selecting machine learning algorithms was to minimize errors in cost forecasting, speed of work, and ease of interpretation of the obtained results: based on which data the decision was made and which data have the most significant influence on the formation of the cost. To minimize the prediction error, detailed data analysis and preparation were carried out for each type of construction vehicle. Many experiments were conducted with various methods for finding and removing anomalous observations and for finding and using the essential features.

In contrast, such methods as Z-index, interquartile range, recursive removal of features, and feature search based on the detection of dependencies using statistical methods were used. A comparative analysis of the results of each of the models was carried out, and the possible reasons for specific results were analyzed. The problems that arise when solving this regression-type problem are presented – the selection of data that best summarizes the system of formation of the cost of a technical vehicle.

Key words: *machine learning, data, algorithm, data processing, machine learning regression models, linear regression, decision tree, random forest.*

Актуальність проблеми. Тема прогнозування вартості транспортних засобів є дуже популярною та піддавалася неодноразовому дослідженню зі сторони науковців та бізнесу. На це є кілька причин – бажання бізнесу передбачати тренди ринку, розуміти формування ціни на вторинному ринку продажу як вживаних транспортних засобів так і нових, тощо. Окрім бізнесу, звичайні громадяни також хочуть мати можливість дізнаватися об'єктивну вартість транспортного засобу, щоб, наприклад, не переплачувати або ж не втратити при продажу власного транспортного засобу. Відповідна тема є цікавою та постійно піддається дослідженню також через те, що економічний фон світу постійно змінюються, а разом з ним змінюються і тренди ринку, в тому числі вторинного ринку транспортних засобів (Massey F. J., 2021; Fushiki T., 2011).

У зв'язку з технічним прогресом та широким застосування техніки та транспортних засобів у найбільш різноманітних сферах та галузях людської діяльності, існує потреба у здійсненні аналізу ринків збуту транспортних засобів та можливості об'єктивної оцінки вартості (Leslie J. R., 2018). Саме тому, обрана тема є актуальною.

Метою дослідження є розробка модулів, що можуть взаємодіяти в інформаційній системі, та які виконують наступні функції: попередній аналіз та обробка структурованих даних про будівельні транспортні засоби різних типів, побудова моделей машинного навчання, зокрема таких як лінійна регресія, дерево прийняття рішень та випадковий ліс, оцінка якості отриманих моделей та продукуючих результатів.

Аналіз останніх досліджень і публікацій. Аналізуючи наукові літературні джерела, було взято до уваги статті та наукові праці, що є схожими до теми дослідження або ж є дотичними до певних аспектів її виконання. Сюди належить – прогнозування вартості легкових транспортних засобів, аналіз транспортних засобів, застосування методів машинного навчання для передбачення ціни, зокрема застосування лінійної, багатовимірної регресії, застосування дерев рішень та використання комбінацій таких методів.

Метою дослідження (Pandey A., 2020; Sharma A. D., 2020) є отримання інформації про найбільш впливові фактори ціноутворення легкових автомобілів. Автори виділяють 5 основних кроків – дослідження ринку та отриманих з нього даних, чистка та підготовка даних, обрання ознак з використанням методу RFE, побудова та оцінка отриманої моделі лінійної

регресії. Значна увага приділяється аналізу кореляцій між залежною ознакою – ціною та незалежними ознаками, а також аналізу кореляції незалежних ознак, які можуть негативно вплинути на остаточну побудовану модель.

У статті (Chen C., 2017; Asghar M., 2021) описано недоліки застосування лінійних моделей для прогнозування цін у зв'язку з тим, що ціна у світі формується за впливу багатьох факторів. Відповідно нелінійні моделі краще прогнозують довільне ціноутворення в реальному житті. Дослідники також пропонують використати S-Curve модель, як альтернативну нелінійну модель для оцінювання вартості вживаних автомобілів.

У статті (Karakoç M. M., 2019; Samruddhi K.) дослідники застосовують лінійну регресію, дерева рішень, та випадковий ліс для прогнозування ціни на авто. Особливість їхнього методу побудови полягає в попередній обробці даних. Автори заповнюють такі пусті значення на основі групування та аналізу даних груп авто, які є схожі на авто з пустим значенням. Для кластеризації груп схожих за характеристиками авто було використано алгоритм кластеризації K-means.

Визначення мети дослідження. Основною метою даної роботи є побудова моделей машинного навчання, яка буде надавати можливість користувачу оцінювати вартість специфічних будівельних транспортних засобів на основі певних технічних та експлуатаційних характеристик.

Виклад основного матеріалу дослідження. Існує багато варіацій алгоритмів методів відбору ознак, які побудовані на основі тренування моделей машинного навчання або ж застосування статистичних методів для статичного аналізу даних.

Дуже важливим етапом, який може значно вплинути на якість моделі є пошук та вилучення викидів. Для боротьби з ними існує ряд статистичних методів, що дозволяють в автоматичному режимі виявити їх (Mammadov H., 2021; Chen C., 2017).

Z-index – ефективний спосіб визначення викидів у наборі даних, якщо набір даних відповідає нормальному розподілу. Z-index кожного спостереження в наборі даних можна обчислити, використовуючи наступну формулу 1:

$$Z = \frac{X - \mu}{\sigma}, \qquad (1)$$

де X – вихідне значення спостереження;

μ – середнє значення нормально розподіленого набору даних;

σ – стандартне відхилення нормально розподіленого набору даних.

Z-index є простим, інтуїтивно зрозумілим та в той же час ефективним методом для виявлення викидів, але його можна застосовувати у випадку нормально розподілених даних. Якщо ж дані не відповідають нормальному розподілу, можна використати ще одним метод, який називається міжквартильний розмах (IQR), який схожий на Z-index та побудована на принципі поділу даних на квантилі.

Модель машинного навчання – це абстракція, яка містить під собою певний алгоритм машинного навчання, що тренується на основі певних даних. Ряд таких алгоритмів є душе широким та може використовувати у своїй роботі різні підходи та різні структури даних.

Лінійна регресія – один з найпростіших алгоритмів машинного навчання, який здатний вирішувати проблему регресії. Побудована модель встановлює залежність між скалярним значенням y та вектором незалежних ознак X. Загальна формула лінійної регресії має наступний вигляд (Формула 2):

$$y = a_0 + a_1 x_1 + a_2 x_2 \ldots + a_n x_n . \qquad (2)$$

Для перевірки адекватності та якості побудованої моделі лінійної регресії достатньо перевірити наступні припущення:

1. Наявність лінійної залежності між залежною та незалежними змінними. Перевірити дану гіпотезу можна наступним чином:

1) Побудувати візуальні графіки діаграми розсіювання та оцінити наявність лінійної залежності;

2) Обрахувати коефіцієнт кореляції Пірсона, який дозволяє оцінити залежність. Коефіцієнт кореляції Пірсона здатний набувати значення в межах від -1 до 1. Абсолютне значення отриманого коефіцієнта кореляції свідчить про значущість лінійної залежності та чим більше це значення – тим сильніша залежність. Знак коефіцієнта кореляції пояснює напрям залежності: позитивний знак означає зростання залежної змінної при зростанні незалежної, від'ємний – спадання залежної змінної при зростанні незалежної.

2. Відсутність мультиколінеарності: незалежні змінні не повинні корелювати між собою, незалежна змінна має мати кореляцію тільки з залежною змінною. Якщо ця гіпотеза не виконуються, це може негативно вплинути на побудовану модель лінійної регресії, оскільки алгоритм не може визначити, яка саме з мультиколінеарних незалежних змінних впливає на формування залежної змінної та з яким ваговим коефіцієнтом.

3. Гомоскедастичність. Це означає, що дисперсія залишків повинна має бути сталою.

4. Значення залишків не повинні корелювати між собою.

Дерево рішень – модель машинного навчання, яка будує дерево специфічної структури для прогнозування. В такій структурі дерева вузол представляє атрибути за якими приймалося рішення, зв'язок – значення атрибуті, на основі яких приймалося рішення від яких залежить цільова функція, листок – значення залежної змінної. Математично дерево рішення можна представити функцією, що приймає на вхід вектор атрибутів $x_1, x_2, \ldots x_n$, а на вихід видає значення y.

Випадковий ліс – один з простих та популярних методів машинного навчання, який використовує ансамблювання. Даний алгоритм працює з використанням bagging принципу. В якості моделі машинного навчання випадковий ліс використовує дерева прийняття рішення. Дерева прийняття рішення мають особливість часто перенавчатися, тому саме техніка ансамблювання має здатність зменшити вплив перенавчання на вихідний результат та покращити якість вихідної моделі.

Універсального правила для підбору оптимальних гіперпараметрів не існує. Найкращим способом щоб це зробити – проведення експериментів з підбору параметрів для мінімізації помилки на тестувальному наборі даних.

Експериментальна частина. Побудова моделі машинного навчання починається з підготовки набору даних, який буде використано для тренування та оцінки результатів. Такий набір даних було попередньо зібрано. Він містить інформацію про технічні та експлуатаційні характеристики міні-вантажників.

Аналіз та побудова моделі машинного навчання розпочинається з загального огляду набору даних (рис. 1).

Набір даних містить 57675 записів спостережень різних моделей міні-вантажників з доволі великим розкидом ціни. Також він містить багато незалежних: як категоріальних так і числових атрибутів, серед яких необхідно виділити ті, які впливають на формування вартості ціни. Для того, щоб відшукати певні залежності між двома атрибутами використовується показник кореляції. Побудуємо матрицю кореляцій для відповідного набору даних (рис. 2).

Аналізуючи отриману матрицю, можна одразу помітити, що атрибут ціни корелює з багатьма іншими атрибутами. Серед основних таких атрибутів міні-вантажника можна виділити: номінальну робочу потужність, вагу, габарити технічного засобу та його вік. Також серед атрибутів є такі важливі характеристики, як максимальна висота підйому та час експлуатації засобу в годинах, проте показник кореляції є малим та не свідчить про якусь залежність.

Дуже важливим для аналізу даних є їх візуальне представлення. Для аналізу залежностей двох атрибутів ідеальним способом представлення даних є побудова точкової діаграми. Точкова діаграма це графік, що відображає значення 2 атрибутів з набору даних у вигляді точок, де значення кожного з атрибутів визначає положення цієї точки. Представимо залежності ціни з іншими атрибутами (рис. 3).

Аналізуючи отримані графіки, одразу можна побачити, що в зібраному наборі даних присутні викиди у значеннях різних атрибутів даних. Тому перш ніж робити певні висновки про важливість атрибутів необхідно позбутися цих викидів та повернутися до подальшого аналізу.

Стратегія пошуку викидів залежить від природи розподілу атрибуту. У випадку нормально розподіленого атрибуту доцільно використовувати алгоритм пошуку та видалення викиду, базуючись на Z-index значенні, інакше – використовувати інший метод, побудований на принципі міжквартильного розкиду.

Рис. 1. Початковий набір даних

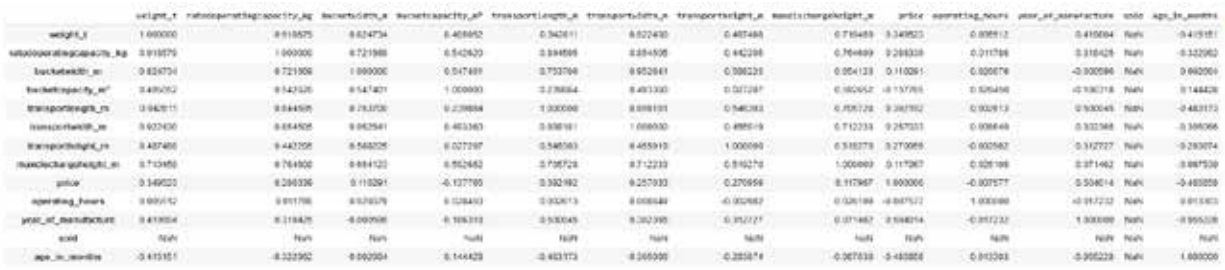

Рис. 2. Матриця кореляцій ознак

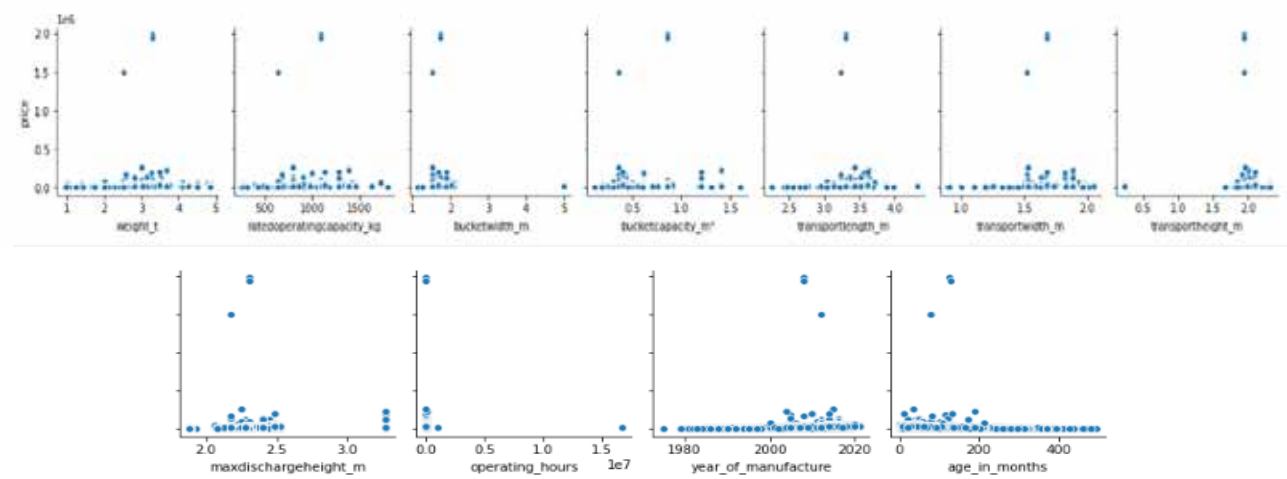

Рис. 3. Візуальне представлення залежності атрибутів з використанням точкової діаграми

Для того щоб визначити, чи дані підлягають під визначення нормального розподілу, існує кілька шляхів. Популярними способами перевірки даних на нормальність є побудова відповідних графіків та візуальна оцінка. Для того щоб визначити нормальність розподілу, можна побудувати наступні діаграми: гістограма розподілу (рис. 4), коробковий графік (рис. 5) або ж Q-Q графіки (рис. 6). При аналізі коробкового графіку також можна виявити наявність викидів (рис. 5).

Це є доволі ефективні способи перевірки розподілу даних на нормальність, які не підлягають автоматизації, а також допускають здійснення помилки оцінювання користувачем, оскільки оцінювання графіків є суб'єктивним.

Для побудови моделі машинного навчання ми використали 3 різні алгоритми, про які згадувалося раніше – лінійну регресію, дерево рішень та випадковий ліс. Саме ці алгоритми є швидкими та в той же час ефективними для вирішення такого типу задач та піддаються легкій інтерпретації

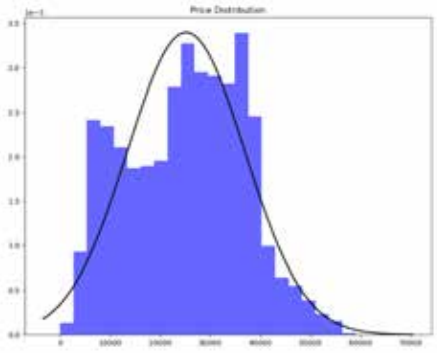

Рис. 4. Гістограма розподілу ціни у порівнянні з теоретичним нормальним розподілом

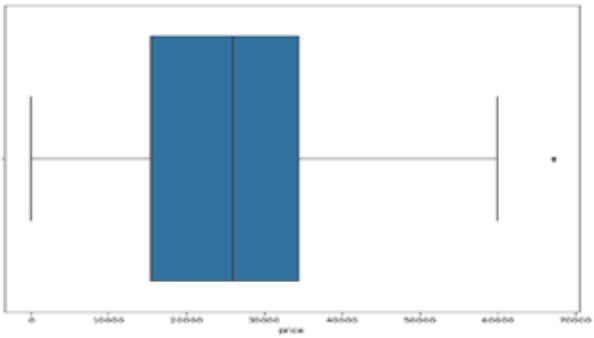

Рис. 5. Коробкова діаграма розподілу ціни

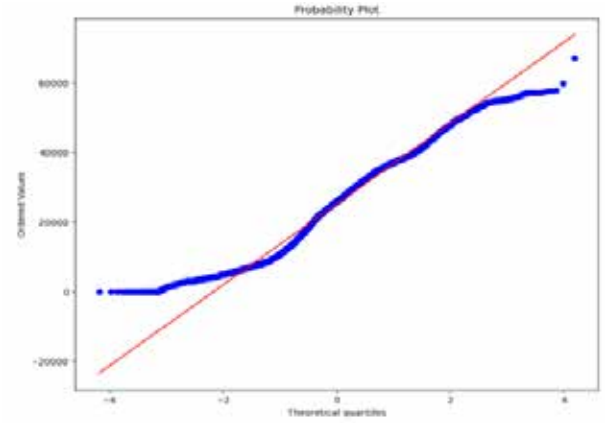

Рис. 6. Q-Q графік розподілу ціни у порівнянні з теоретичним нормальним розподілом

результатів. Проте, перш ніж будувати модель необхідно визначити, які саме атрибути набору даних будуть використовуватися в процесі тренування та будуть найбільш ефективними. Для цього існують алгоритми відбору ознак.

В якості алгоритму відбору ознак слід використати алгоритм рекурсивного видалення ознак, що належать до обгорткових методів відбору ознак. В якості моделі, що буде використовуватися для відбору ознак, виберемо модель дерева рішень.

Отож, будемо будувати послідовно кілька моделей машинного навчання одного типу, видаляючи при цьому по одній найменш впливовій ознаці з вхідного набору даних.

Для об'єктивної оцінки якості побудованої моделі машинного навчання було вирішено використовувати алгоритм k-кратного перехресного затверджування [4]. Даний алгоритм розділяє первинний вхідний набір даних на k піднаборів однакової розмірності. З k піднаборів даних алгоритм обирає один та використовує його як тестувальний набір для оцінки натренованої моделі, а інші k-1 піднаборів – використовуються в етапі тренування моделі машинного навчання. Згодом процес перехресного затверджування повторюється ще k разів, при цьому обираючи один з піднаборів даних як тестувальний, а інші – як тренувальні. Таким чином ми отримує k оціночних результатів тренування-тестування моделі, які можемо

Таблиця 1

Результати тренування та оцінки лінійної регресії

Features	Train				Test			
	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
Підйомна сила	6293.212 +- 9.253	65442848.551 +- 184877.439	8089.667 +- 11.427	0.070 +- 0.001	6293.645 +- 73.149	65460029.696 +- 1663143.458	8090.084 +- 102.799	0.070 +- 0.013
Підйомна сила Вік транспортного засобу	5465.657 +- 7.957	46531041.425 +- 111916.565	6821.362 +- 8.205	0.339 +- 0.002	5466.116 +- 70.216	46539009.562 +- 1006614.740	6821.553 +- 73.662	0.338 +- 0.015
Підйомна сила Вік транспортного засобу Експлуатаційні години	5465.486 +- 7.976	46526121.105 +- 112374.35	6821.001 +- 8.239	0.339 +- 0.002	5466.118 +- 70.304	46536284.822 +- 1011098.339	6821.350 +- 73.983	0.339 +- 0.015
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага	5433.364 +- 7.973	46218350.414 +- 111012.505	6798.403 +- 8.167	0.343 +- 0.002	5434.160 +- 71.160	46230035.460 +- 999161.210	6798.872 +- 73.324	0.343 +- 0.014
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага Номінальна потужність	4712.154 +- 6.823	35921980.793 +- 106659.796	5993.488 +- 8.905	0.490 +- 0.002	4713.033 +- 59.820	35934788.235 +- 959968.876	5994.036 +- 79.533	0.489 +-0.014
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага Номінальна потужність Максимальна підйомна висота	4712.149 +- 6.821	35921791.950 +- 106659.224	5993.473 +-8.905	0.490 +- 0.002	4713.175 +- 59.825	35936468.463 +- 959769.522	5994.176 +- 79.516	0.489 +- 0.014
Підйомна сила Вік транспортного засобу Експлуатаційні характеристика Вага Номінальна потужність Максимальна підйомна висота Об'єм ковша	4595.790 +- 6.675	34858125.505 +- 102213.394	5904.071 +- 8.661	0.505 +- 0.001	4596.971 +- 58.895	34875773.531 +- 919794.691	5905.062 +- 77.547	0.504 +- 0.013

в подальшому агрегувати. Такий підхід дозволяє уникнути помилкових оцінок моделі, яка натренувалася на “щасливому” наборі даних, який показує найкращий результат і дає нам можливість усереднити результат на сукупній вибірці.

Отож було проведення тренування та оцінка вищезгаданих алгоритмів машинного навчання на наборі даних міні-вантажників. Відповідні результати представлені у вигляді таблиць нижче: лінійна регресія (табл. 1), дерево рішень (табл. 2), випадковий ліс (табл. 3).

Аналізуючи отримані результати можна зробити висновок, що випадковий ліс впорався з поставленою задачею найкраще та показує доволі хороші результати.

Таким чином випадковий ліс показує середню абсолютну похибку 3000 на тестувальному та 2100 на тренувальному наборі даних (табл. 3). Коефіцієнт детермінації також є хорошим – 0.79 та 0.7 для тренувальної та тестувальної вибірки, що означає, що дана модель є адекватною та доволі якісно описує загальну вибірку даних.

Дерево рішень – показує посередні результати. 3200 та 1900 на тестувальній та тренувальній вибірці, а коефіцієнт детермінації – 0.8 та 0.65 відповідно (табл. 2). Також за результатами тренування та оцінки якості моделі можна помітити, що модель дерева рішень піддається перенавчанню на тренувальній вибірці. Це власне і є основна проблема даного алгоритму.

Лінійна регресія показала найменш точні результати серед обраних та тренованих моделей машинного навчання. Її показники були наступними – 4595 та 4596 для тренувальної та тестувальної вибірки, а коефіцієнт детермінації склав 0.5 для обидвох (табл. 1). Таким чином лінійна регресія не піддалася перенавчанню на тренувальному наборі даних, але сама не дозволяє описати природу залежностей відповідного набору даних. Спостереження вибірки є доволі розкиданими та можливо місять певні нелінійні залежності, які лінійна регресія задовільнити не може.

Окрім набору даних, що містить інформацію про міні-вантажники, було також зібрано вибірку з інформацією про міні-екскаватори.

Таблиця 2

Результати тренування та оцінки дерева рішень

Features	Train				Test			
	MAE	MSE	RMSE	R2	MAE	MSE	RMSE	R2
Підйомна сила	5370.066 +- 41.205	45106510.157 +- 813098.940	6715.867 +- 60.309	0.359 +- 0.011	5371.452 +- 82.583	45135372.166 +- 1264421.932	6717.631 +- 93.861	0.358 +- 0.021
Підйомна сила Вік транспортного засобу	4139.388 +- 29.796	29205102.875 +- 283522.107	5404.111 +- 26.191	0.585 +- 0.004	4220.810 +- 62.135	30122660.667 +- 948910.096	5487.741 +- 85.812	0.572 +- 0.012
Підйомна сила Вік транспортного засобу Експлуатаційні години	1991.326 +- 22.793	13866846.254 +- 132301.988	3723.779 +- 17.740	0.803 +- 0.002	0.004 +- 67.973	24627498.323 +- 1028281.471	4961.530 +- 103.526	0.650 +- 0.015
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага	1977.431 +- 5.645	13796859.235 +- 66602.690	3714.402 +- 8.976	0.804 +- 0.001	3174.007 +- 61.837	24532473.849 +- 974267.995	4952.047 +- 98.504	0.651 +-0.014
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага Номінальна потужність	1977.431 +- 5.645	13796859.235 +- 66602.690	3714.402 +- 8.976	0.804 +- 0.001	3172.893 +- 59.523	24534251.369 +- 934127.440	4952.307 +- 94.400	0.651 +-0.013
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага Номінальна потужність Максимальна підйомна висота	1977.431 +- 5.645	13796859.235 +- 66602.690	3714.402 +- 8.976	0.804 +- 0.001	3174.043 +- 58.777	24565235.683 +-935534.464	4955.430 +- 94.586	0.651 +- 0.014
Підйомна сила Вік транспортного засобу Експлуатаційні характеристика Вага Номінальна потужність Максимальна підйомна висота Об’єм ковша	1977.431 +- 5.645	13796859.235 +- 66602.690	3714.402 +- 8.976	0.804 +- 0.001	3174.375 +- 57.040	24549672.483 +- 920909.236	4953.889 +- 93.043	0.651 +- 0.014

Таблиця 3

Результати тренування та оцінки випадкового лісу

Features	Train				Test			
	MAE	**MSE**	**RMSE**	**R2**	**MAE**	**MSE**	**RMSE**	**R2**
Підйомна сила	5378.565 +- 45.176	45275751.220 +- 875025.68	6728.413 +- 64.904	0.357 +- 0.013	5382.977 +- 76.338	45347582.833 +- 1213065.022	6733.460 +- 90.014	0.355 +- 0.016
Підйомна сила Вік транспортного засобу	4137.263 +- 26.598	29158564.515 +- 270105.054	5399.810 +- 24.935	0.586 +- 0.004	4214.393 +- 58.268	30039185.418 +- 900312.558	5480.198 +- 81.323	0.573 +- 0.012
Підйомна сила Вік транспортного засобу Експлуатаційні години	2357.318 +- 17.910	14737814.210 +- 107742.117	3838.961 +- 14.010	0.791 +- 0.001	3090.553 +- 54.758	21224145.439 +- 749433.857	4606.256 +- 80.911	0.698 +- 0.011
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага	2351.151 +- 5.783	14707156.196 +- 66646.522	3834.981 +- 8.697	0.791 +- 0.001	3081.015 +- 57.240	21153848.877 +- 804216.659	4598.508 +- 87.033	0.699 +- 0.012
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага Номінальна потужність	2351.265 +- 5.798	14708204.877 +- 67275.425	3835.118 +- 8.780	0.791 +- 0.001	3082.581 +- 56.649	21159954.561 +- 796869.75	4599.187 +- 86.233	0.699 +- 0.012
Підйомна сила Вік транспортного засобу Експлуатаційні години Вага Номінальна потужність Максимальна підйомна висота	2351.178 +- 5.472	14708979.438 +- 66814.503	3835.219 +- 8.719	0.791 +- 0.001	3081.406 +- 56.880	21153321.615 +- 791325.308	4598.476 +- 85.665	0.699 +- 0.011
Підйомна сила Вік транспортного засобу Експлуатаційні характеристика Вага Номінальна потужність Максимальна підйомна висота Об'єм ковша	2351.026 +- 5.996	14708228.258 +- 68283.570	3835.120 +- 8.912	0.791 +- 0.001	3081.759 +- 54.814	21151557.368 +- 781739.460	4598.303 +- 84.656	0.699 +- 0.011

Таблиця 4

Найкращі результати тренування та оцінки моделей машинного навчання для набору даних міні-екскаваторів

Модель	Train				Test			
	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
Лінійна регресія	4620.016 +- 19.371	39700909.275 +- 360633.978	6300.801 +- 28.636	0.574 +- 0.003	4626.739 +- 169.046	39812037.807 +- 3243015.909	6304.476 +- 256.170	0.572 +- 0.026
Дерево рішень	235.933 +- 7.434	913381.498 +- 52533.676	955.310 +- 27.660	0.990 +- 0.001	3549.773 +- 256.752	41144201.893 +- 6822245.189	6392.429 +- 530.139	0.557 +- 0.071
Випадковий ліс	1238.417 +- 13.583	4194736.099 +- 111487.911	2047.926 +- 27.133	0.915 +- 0.001	3070.881 +- 215.310	25501721.283 +- 4138073.619	5033.298 +- 409.434	0.726 +- 0.039

Цей набір даних був значно менший, аніж міні-вантажники та містив 4038 спостережень після видалення викидів та обробки пустих значень. Він також містив доволі багато атрибутів, проте після аналізу залишились тільки найважливіші – вага, об'єм ковша, ширина бази, максимальна, глибина копання, сила копання, потужність двигуна, кількість експлуатаційних годин та рік виробництва.

Після проведення тренування та оцінки отриманий моделей машинного навчання, були отримані такі найкращі результати (табл. 4).

Отримані результати є доволі схожими на попередні. Лінійна регресія показала найгірші результати з коефіцієнтом детермінації 0.572. Це означає, що лінійна регресія не здатна описати природу походження та залежностей даних.

Аналізуючи тренувальні та тестувальні оцінки дерева рішень, можна зробити висновок, що під час тренування модель піддалася сильному перенавчанню, оскільки тренувальні коефіцієнт детермінації є надзвичайно високим, а похибки є надзвичайно малими, а тестувальні навпаки. Таку модель використовувати не можна, оскільки вона не узагальнює існуючі залежності та відповідно буде давати помилкові результати.

Випадковий ліс – оптимальна серед трьох натренованих моделей, оскільки результати на тестувальних вибірках в середньому становлять 3000 абсолютної похибки, а коефіцієнт детермінації – 0.726, що означає, що модель достатньо добре пояснює існуючі залежності. Тренувальна оцінка є значно кращою, ніж тестувальна, що також свідчить про наявність перенавчання.

Метою відбору алгоритмів машинного навчання була мінімізація похибок при прогнозуванні вартості, швидкість роботи, легкість інтерпретації отриманих результатів: на основі яких даних приймалося рішення та які дані найбільше впливають на формування вартості. Тому вибрано та проведено експерименти з використанням трьох моделей – лінійної регресії, дерева рішень та випадкового лісу, останній з яких показав найбільш точні та стабільні результати.

Висновки. Також для мінімізації похибки прогнозування було проведено детальний аналіз даних та їх підготовку для кожного типу будівельного транспортного засобу. Такі дії важко піддаються автоматизацію та вимагають попереднього аналізу людиною, щоб мінімізувати фактори, що можуть негативно повпливати на подальші результати та уникнути видалення реальних спостережень з набору даних. Було проведено багато експериментів з різними методами для пошуку та видалення аномальних спостережень, для пошуку та використання найбільш важливих ознак, при цьому використовувалися такі методи як Z-index, міжквартильний розмах, рекурсивне видалення ознак, пошук ознак на основі виявлення залежностей з використанням статистичних методів, тощо.

Фінальні результати, що були отримані в результаті експериментів у проведенні аналізу даних та побудови регресійних моделей, були доволі точними – середня абсолютна похибка для різних наборів даних склала близько 3000, що є хорошим показником та може бути пояснена відповідним розкидом цін на ринку. Було проведено порівняльний аналіз результатів кожної з моделей, проаналізовано можливі причини тих чи інших результатів.

Основна проблема при вирішенні даної задачі регресійного типу – відбір даних, що якнайкраще узагальнюють систему формування вартості технічного транспортного засобу. Оскільки дані були зібрані з реальних платформ розміщення оголошень про продаж таких транспортних засобів, відповідно прогнозований результат може бути з певною похибкою, оскільки присутня суб'єктивна оцінка вартості транспортних засобів їх власником.

ЛІТЕРАТУРА:

1. Massey F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association. 2021. № 46 (253).* P. 68–78. DOI: 10.1080/01621459.1951.10500769.

2. Leslie J. R., Stephens M. A., Fotopoulos S. Asymptotic Distribution of the Shapiro-Wilk W for Testing for Normality. *Ann. Statist.2018. № 14(4).* P. 214-224. DOI: 10.1214/aos/1176350172.

3. Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput. 2011. № 21(2)*. P. 137–146. DOI: 10.1007/s11222-009-9153-8.

4. Mammadov H. Car Price Prediction in the USA by using Liner Regression. *International Journal of Economic Behavior (IJEB). 2021. № 11(1).* P. 56-68. DOI: 10.14276/2285-0430.3049.

5. Pandey A., Rastogi V., Singh S. Car's Selling Price Prediction using Random Forest Machine Learning Algorithm. *SSRN Journal. 2020. № 1.* P. 146-159. DOI: 10.2139/ssrn.3702236.

6. Fadzilah S. Nur A. A. Used Car Price Estimation: Moving from Linear Regression towards a New S-Curve Model. *IJBS.2021. № 22(3).* P. 1174–1187. DOI: 10.33736/ijbs.4293.2021.

7. Chen C., Hao L., Xu C. Comparative analysis of used car price evaluation models. *Hangzhou. 2017. № 1.* P.201-210. DOI: 10.1063/1.4982530.

8. Sharma A. D., Sharma V., Mittal S., Jain G., Narang S. Predictive analysis of used car prices using machine learning. *International Research Journal of Modernization in Engineering Technology and Science. 2020. № 3(6).* P. 11-20.

9. Chen Y., Li C., Xu M. Business Analytics for Used Car Price Prediction with Statistical Models. *3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021), Guangzhou, China, 2021.* P. 20-32. DOI: 10.2991/assehr.k.211209.090.

10. Karakoç M. M., ÇeliK G., Varol A. Car Price Prediction Using An Artificial Neural Network. *2019. № 2*. P. 5-19.

11. Samruddhi K., Ashok Kumar R. Used Car Price Prediction using K-Nearest Neighbor Based Model. *International Journal of Innovative Research in Applied Sciences and Engineering. 2020. № 4(3)*. P. 686–689. DOI: 10.29027/IJIRASE.v4.i3.2020.686-689.

12. Asghar M., Mehmood K., Yasin S., and Khan Z., Used Cars Price Prediction using Machine Learning with Optimal Features. *Pakistan Journal of Engineering and Technology*. *2021. vol. 4, no. 2*. P. 113-119.

REFERENCES:

1. Massey F. J. (2021). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association. № 46 (253).* P. 68–78. DOI: 10.1080/01621459.1951.10500769.

2. Leslie J. R., Stephens M. A., Fotopoulos S. (2018). Asymptotic Distribution of the Shapiro-Wilk W for Testing for Normality. *Ann. Statist. № 14(4).* P. 214-224. DOI: 10.1214/aos/1176350172.

3. Fushiki T. (2011). Estimation of prediction error by using K-fold cross-validation. *Stat Comput. № 21(2)*. P. 137–146. DOI: 10.1007/s11222-009-9153-8.

4. Mammadov H. (2021). Car Price Prediction in the USA by using Liner Regression. *International Journal of Economic Behavior (IJEB). № 11(1).* P. 56-68. DOI: 10.14276/2285-0430.3049.

5. Pandey A., Rastogi V., Singh S. (2020). Car's Selling Price Prediction using Random Forest Machine Learning Algorithm. *SSRN Journal. № 1.* P. 146-159. DOI: 10.2139/ssrn.3702236.

6. Fadzilah S. Nur A. A. (2021). Used Car Price Estimation: Moving from Linear Regression towards a New S-Curve Model. *IJBS. № 22(3).* P. 1174–1187. DOI: 10.33736/ijbs.4293.2021.

7. Chen C., Hao L., Xu C. (2017). Comparative analysis of used car price evaluation models. *Hangzhou. № 1.* P.201-210. DOI: 10.1063/1.4982530.

8. Sharma A. D., Sharma V., Mittal S., Jain G., Narang S. (2020). Predictive analysis of used car prices using machine learning. *International Research Journal of Modernization in Engineering Technology and Science. № 3(6).* P. 11-20.

9. Chen Y., Li C., Xu M. (2021). Business Analytics for Used Car Price Prediction with Statistical Models. *3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021), Guangzhou, China.* P. 20-32. DOI: 10.2991/assehr.k.211209.090.

10. Karakoç M. M., ÇeliK G., Varol A. (2019). Car Price Prediction Using An Artificial Neural Network. *№ 2*. P. 5-19.

11. Samruddhi K., Ashok Kumar R. (2020). Used Car Price Prediction using K-Nearest Neighbor Based Model. *International Journal of Innovative Research in Applied Sciences and Engineering. № 4(3).* P. 686–689. DOI: 10.29027/IJIRASE.v4.i3.2020.686-689.

12. Asghar M., Mehmood K., Yasin S., and Khan Z. (2021). Used Cars Price Prediction using Machine Learning with Optimal Features. *Pakistan Journal of Engineering and Technology*. *vol. 4, no. 2*. P. 113-119.