УДК 004.42
DOI https://doi.org/10.32782/2663-5941/2023.1/18

Марчук Д.К.
Державний університет «Житомирська політехніка»

Кравченко С.М.
Державний університет «Житомирська політехніка»

Левченко А.Ю.
Державний університет «Житомирська політехніка»

Лежньов І.Я.
Державний університет «Житомирська політехніка»

ВИКОРИСТАННЯ ЧАСОВИХ РЯДІВ ПРИ ПРОГНОЗУВАННІ ГРОШОВОЇ ВАРТОСТІ АВТОМОБІЛІВ

У статті розглянуто задачу прогнозування грошової вартості автомобілів за параметрами з використанням часових рядів. В даний час завдання прогнозування є актуальним для різного практичного застосування. Це можуть бути такі сфери діяльності як економіка, фінанси, бізнес, торгівля, політика тощо. Прогнозування часового ряду вирішується на основі створеної моделі, яка описує досліджуваний процес. Об'єктом дослідження є застосування моделей авторегресійної інтегрованої ковзної середньої (ARIMA) для прогнозування вартості автомобіля з використанням часових рядів. Моделі ARIMA здатні моделювати широкий спектр сезонних даних. В процесі дослідження створено власний датасет за допомогою онлайн сервісу, який складався із 10 000 записів, дані якого збережені у форматі CSV. В якості інструменту реалізації було обрано мову програмування Python, а саме бібліотеки: pandas, numpy, matplotlib, seaborn, statsmodels, itertools і keras. Результати дослідження представлені візуально. Було проведено аналіз за різними параметрами: вартістю автомобіля, роком випуску, видом коробки передач, типом палива. Аналіз проведено різними засобами та методами ARIMA. В результаті аналізу можна зробити висновок, що обрана модель задовольняє умовам прогнозування часових рядів. Аналізуючи графіки можна побачити сезонність та стрімкий ріст вартості автомобілів на майбутні роки, причому щорічно буде певний спад ціни в певному сезоні кварталу. За отриманими результатами сезонності та тенденцій можна зробити висновок, що ціна автомобіля має сезонність зростання ціни, що дає можливість спрогнозувати свої дії та прийняти зважене рішення щодо купівлі або продажу автомобіля.

Ключові слова: *автомобіль, прогнозування, часовий ряд, ARIMA, тренд, сезонність.*

Постановка проблеми. Щоденно відбувається безліч купівель та продажів. Серед них не є винятком автомобіль як об'єкт продажі та купівлі. При купівлі автомобіля людина звертає увагу на низку факторів, таких як: пробіг, комплектація, тип двигуна, тип палива, об'єм, рік випуску тощо.

Для того щоб купити автомобіль також потрібно розуміти за яку ціну його можна буде купити з плином часу чи продати через певний час. Саме тому потрібно розуміти майбутню ринкову ціну. Для цього можна використовувати інструменти прогнозування.

Актуальність теми полягає в необхідності прийняття відповідного рішення, щодо купівлі чи продажу автомобіля. Знаючи вартість на поточний момент та ймовірну вартість на ринку у майбутньому людина може прийняти рішення про продаж автомобіля зараз чи відкласти це питання на потім, коли вартість автомобіля буде вигідна порівняно з поточною чи відкласти продаж.

Аналіз останніх досліджень і публікацій. Метою аналізу даних є виявлення корисної інформації, здійснення висновків і прийняття рішень. Аналіз даних може здійснюватися різними методами математичними, статистичними, інтелектуальними, візуальними.

У статті [1] досліджуються алгоритми інтелектуального аналізу даних, які на основі правил і обчислень дозволяють створити модель, що аналізує дані, здійснюючи пошук певних закономірностей і тенденцій. Шляхом дослідження алгоритмів інтелектуального аналізу даних було

розроблено моделі та методи для встановлення впливу одних хронічних захворювань на інші. Проведені дослідження свідчать про перспективність використання методів інтелектуального аналізу даних для підвищення якості медичної допомоги пацієнтам.

У статті [2] описано новий підхід до використання інтелектуальних технологій для певних бізнес-рішень, а саме для дослідження цінової політики вартості будинків залежно від їх розмірів. У статті розглядається метод математичного програмування, а саме метод градієнтного спуску.

У статті [3] описано дослідження можливостей застосування платформи ML.NET для прогнозування оцінки кредитоспроможності фізичних осіб. Метою роботи є розробка системи, яка на основі проведеного аналізу даних визначатиме кредитоспроможність фізичних осіб.

У роботі [4] авторами продемонстровано програмний продукт для проведення аналізу кадрів відеоданих з метою пошуку вільного місця на парковці.

При прогнозуванні часто застосовують моделі авторегресійної інтегрованої ковзної середньої (ARIMA). Наприклад для оцінки цін на зерно автори в статті [5] використовували методи авторегресійної інтегрованої ковзної середньої, а точність прогнозів перевіряли з використанням стандартних стандартів середньої квадратичної помилки і середньої абсолютної помилки у відсотках. Моделі ефективності ARIMA як інструмент прогнозування цін були ефективно продемонстровані реалістичними моделями прогнозованих цін на 2020 рік.

У дослідженнях [6-9] модель авторегресійної інтегрованої ковзної середньої використовувалася для аналізу часової динаміки поширення COVID-19 на основі попередньо зібраних статистичних даних.

На основі даних часових рядів зібраних з 1978 року по 2017, автори статті [10] зробили прогноз на витрати на охорону здоров'я в Китаї з 2018 по 2022 рік, а також показали теоретичні основи для коригування політики охорони здоров'я.

Постановка завдання. Метою роботи є дослідження особливостей інтелектуального аналізу даних та реалізації прогнозування вартості автомобіля. Встановлена мета обумовлює наступні завдання:

– формування датасету;

– вибір моделей для прогнозування;

– реалізація прогнозування вартості автомобіля за різними ознаками;

– аналіз результатів.

Виклад основного матеріалу дослідження. Якісний, кількісний, системний, комбінований це основні підходи для прогнозування. Деякі базуються на математичних моделях, деякі на історичних даних, деякі на досвіді або інтуїції експертів. Але важливо правильно зробити оцінку отриманому прогнозу. Для оцінки якості прогнозу існує багато методів, до яких можна віднести Mean Average Deviation, Running Sum of Forecast Error або Tracking Signal.

Для вирішення поставленого завдання доцільно використовувати часові ряди, тобто це послідовність даних дискретного часу. Зазвичай часові ряди представляються лінійними графіками. Прогнозування за часовими рядами – це використання різних моделей, які можуть зробити прогноз за даними попередніх періодів. Популярними і широко застосовуваними статистичними методами прогнозування часових рядів є моделі ARIMA (англ. Autoregressive integrated moving average). Моделі ARIMA націлені на опис автокореляцій у даних, які розглядаються. Існують сезонні та несезонні моделі.

Використовується стандартне позначення ARIMA(p, d, q), де параметри замінюються цілочисельними значеннями для швидкої вказівки конкретної використовуваної моделі ARIMA.

Параметри моделі ARIMA визначаються наступним чином:

1. p - число спостережень відставання, включених в модель, такзване порядком відставання.

2. d - кількість разів, коли вихідні спостереження розрізняються, також називається ступенем відмінності.

3. q - розмір вікна ковзної середньої, також званий порядком ковзної середньої.

У загальному вигляді модель ARMA (p, q), де

p - порядок авторегресії,

q - порядок змінного середнього, виглядає наступним чином:

$$y'_t = a_1 y_{t-1} + \ldots + a_p y_{p-t} + \varepsilon_t + \theta_1 y_{1-t} + \ldots + \theta_q y_{t-q}$$

Значення 0 може бути використано для параметра, який вказує, що цей елемент моделі не використовується. Таким чином, модель ARIMA може бути налаштована для виконання функції моделі ARMA і навіть простої моделі AR, і / або MA.

Ухвалення моделі ARIMA для тимчасового ряду передбачає, що базовий процес, який справив спостереження, є процесом ARIMA. Це може здатися очевидним, але допомагає мотивувати необхідність підтвердження припущень моделі в необроблених спостереженнях і в залишкових помилки прогнозів з моделі.

Для проведення аналізу даних було сформовано датасет із використанням онлайн сервісу [11] (Таблиця 1).

Таблиця 1

Параметри датасету

№	Назва	Тип даних	Опис
1	Id	Int (index)	Індекс айди
2	Name	String	Назва авто
3	Fuel_type	String(choocie)	Тип палива
4	Engaine_liter	Float	об'єм двигуна
5	horse_power	Int	Кінські сили
6	Kpp	String(choocie)	Тип коробки передач
7	Body_style	String(choocie)	Тип кузова
8	Drive_ wheels	String(choocie)	Тип приводу
9	Color	String(choocie)	Колір
10	Specification	String(choocie)	Комплектація
11	Year_vupusk	Date	Дата випуску
12	Price	float	Ціна
13	probeg	int	Пробіг

Генерування даних було із можливістю збереження у CSV-форматі. Результатом став файл датасету із генерованими даними, де знаходиться 10 000 записів, які у часовому розрізі являють собою останні тридцять років.

Для первинного аналізу дата сету було використано описову статистику(рис. 1).

Описова статистика включає ті функції, що підсумовують центральну тенденцію, дисперсію та форму розподілу набору даних, виключаючи NaN значення.

Для аналізу зв'язку між змінними використана функція кореляції, дані якої представлені на рисунку 2.

Для подальшого аналізу всі дані було відфільтровано та залишено ключові поля для time series – дату та значення. Обрано було параметрами тип палива та комплектація, тому що за аналізом датасету це є важливі характеристики при купівлі продажу автомобіля.

Часові ряди мають очевидну сезонність, а також загальну тенденцію до зростання. Можна візуалізувати дані за допомогою методу, який називається seasonal_decompose – розкладанням часових рядів. Часовий ряд можна розкласти на три різні компоненти: тенденцію, сезонність та шум.

Головне при підборі даних часових рядів в сезонної моделі ARIMA – знайти значення ARIMA (p, d, q) (P, D, Q) s, які оптимізують необхідний показник. Для кожної комбінації параметрів функція SARIMAX () з модуля statsmodels може підібрати нову сезонну модель ARIMA і оцінити її загальну якість. Оптимальним набором параметрів буде той, в якому потрібні критерії найбільш продуктивні. Для початку згенеруємо різні комбінації параметрів:

	id	name	fuel_type	engaine_liter	horse_power	kpp	body_style	drive_wheels	color	specification	year_vupusk	price	probeg
0	100	nissan	gas	4.4	148	automatic	coupe	4wd	none	indevidual	2019-03-11	109964.40	138861
1	101	saab	gas	2.9	601	headdrive	universal	bwd	white	luxaary	1993-09-17	20966.05	54162
2	102	bmw	gas	3.8	114	automatic	sedan	bwd	light-grey	luxaary	1994-04-08	184429.63	111976
3	103	bmw	gas	1.1	626	automatic	coupe	bwd	yellow	avrage	2017-06-09	286730.29	90717
4	104	mazda	disel	1.0	512	headdrive	cabriolet	bwd	blue	indevidual	1999-01-04	157914.11	55381

Рис. 1. Дані датасету

	id	name	fuel_type	engaine_liter	horse_power	kpp	body_style	drive_wheels	color	specification	year_vupusk	price	probeg
0	1.0	5831.5	7510.5	6857.0	914.5	2462.5	3009.5	1598.5	4917.5	5797.5	9740.0	3634.0	9132.0
1	2.0	7427.5	7510.5	3872.0	6891.5	7462.5	9002.5	4889.0	8609.5	7465.0	1180.0	623.0	2631.0
2	3.0	1436.0	7510.5	5656.5	462.5	2462.5	7004.5	4889.0	4003.0	7465.0	1384.5	6152.0	7017.0
3	4.0	1436.0	7510.5	222.5	7224.0	2462.5	3009.5	4889.0	9542.0	834.5	9175.0	9562.0	5455.0
4	5.0	4615.0	2510.5	57.5	5719.5	7462.5	1021.0	4889.0	2198.5	5797.5	2963.0	5247.0	2715.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996.0	5831.5	7510.5	2440.5	8861.0	2462.5	7004.5	8291.0	7961.0	834.5	9817.0	1266.0	5954.0
9996	9997.0	2924.5	7510.5	834.5	8065.0	7462.5	1021.0	8291.0	6745.0	4111.0	6485.0	7392.0	7908.0
9997	9998.0	6637.0	7510.5	222.5	2697.0	7462.5	1021.0	4889.0	1330.0	834.5	7446.0	4076.0	8606.0
9998	9999.0	2824.5	7510.5	3852.5	3540.5	2462.5	9002.5	4889.0	2198.5	7465.0	3680.5	7862.0	3546.0
9999	10000.0	8197.0	7510.5	2834.5	4728.5	2462.5	4991.0	8291.0	2198.5	834.5	2687.0	8895.0	9571.0

Рис. 2. Результат кореляція

```
p = d = q = range(0, 2)
pdq = list(itertools.product(p, d, q))
seasonal_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.pr
oduct(p, d, q))]

print('Приклади комбінацій параметрів для сезонного ARIMA ...')
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[1]))
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[2]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[3]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[4]))
```

Тепер можна використовувати певні триплети параметрів для автоматизації процесу оцінки моделей ARIMA по різним комбінаціям. При оцінці і порівнянні статистичних моделей, що відповідають різним параметрам, враховується, наскільки та чи інша модель відповідає даним і наскільки точно вона здатна прогнозувати майбутні точки даних.

Використаємо значення AIC (Akaike Information Criterion), які підходять для роботи з моделями ARIMA на основі statsmodels. AIC оцінює, наскільки добре модель відповідає даним, беручи до уваги загальну складність моделі.

Чим менше функцій використовує модель, щоб досягти відповідності даним, тим вище її показник AIC. Тому потрібно знайти модель з найменшим значенням AIC (рис. 3).

Відповідно до отриманих даних, SARIMAX (1, 1, 1) x (0, 1, 1, 12) отримує найменший показник AIC (7819.115). Отже, ці параметри можна вважати оптимальними. Цю модель можна проаналізувати більш детально. Додаємо оптимальні параметри в модель SARIMAX. Для отримання інформації про модель можна скористатися методом summary. Атрибут summary повертає багато інформації, але треба зосередитися на таблиці коефіцієнтів (рис. 4).

При підборі сезонних моделей ARIMA важливо проводити діагностику моделі, щоб переконатися, що жодне з припущень, зроблених моделлю, не було порушено.

Для початку потрібно порівняти прогнозовані значення з реальними значеннями часового ряду, що допоможе зрозуміти точність прогнозів. Про-

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0358	0.154	-0.233	0.816	-0.337	0.265
ma.L1	-1.1473	0.020	-58.250	0.000	-1.186	-1.109
ar.S.L12	-0.0396	0.194	-0.204	0.838	-0.420	0.341
ma.S.L12	-0.8107	0.100	-8.072	0.000	-1.007	-0.614
sigma2	1.021e+10	1.14e-12	8.97e+21	0.000	1.02e+10	1.02e+10

Рис. 3. Комбінації параметрів

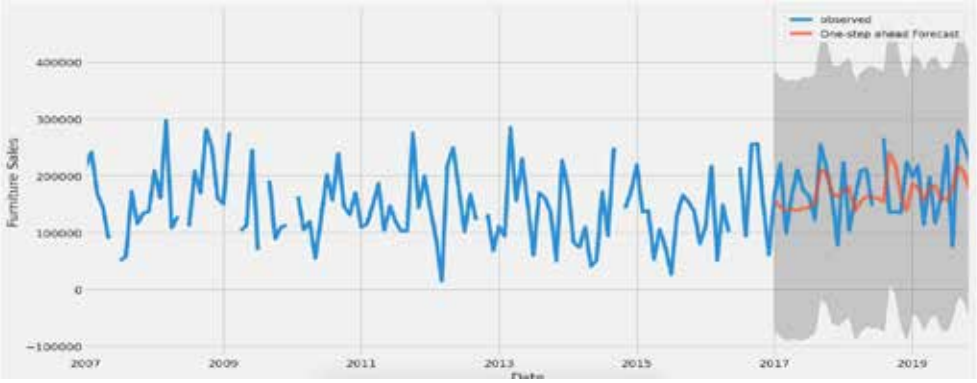

Рис. 4. Результат метода summary

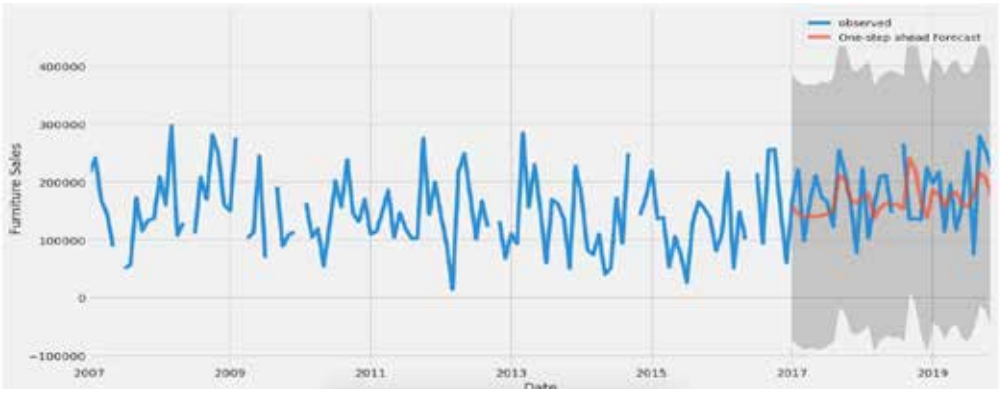

Рис. 5. Графік прогнозування значень з реальними значеннями

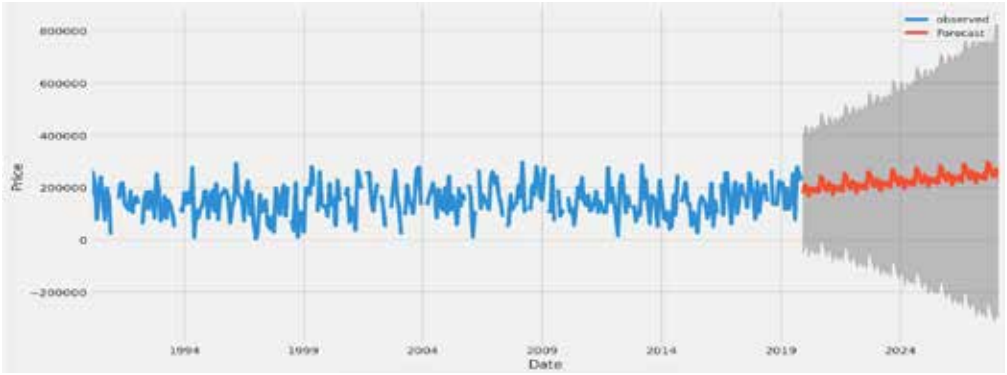

Рис. 6. Графік прогнозування

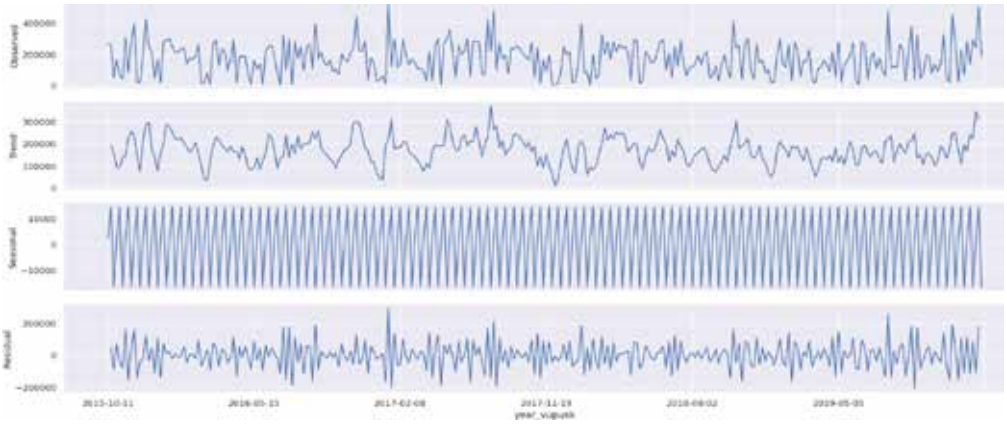

Рис. 7. Графік трендів та сезонності

гнозування почнеться із 2017 року. Буде використано покрокове прогнозування, а це означає, що прогнози в кожній точці генеруються з використанням повної історії аж до цієї точки (рис. 5).

Тепер можна використовувати модель ARIMA для прогнозування майбутніх значень. Отримані дані(рис.6) показують, що тимчасові ряди маю свою характерну тенденцію росту та спаду, що дає розуміння певного шаблону коливання ціни на авто, що вказує на періодичне продовжуване стабільне зростання, що повторюватиметься час від часу.

Результати. Для порівняльного аналізу були взяті дані по автомобілях що мають однаковий тип коробки передач та тип палива. Було отримано їх сезонність та трендовість. Результати порівняння наведено на рисунку 7.

Схожі дані були отримані при прогнозуванні (рис. 8).

Графік описує тренд, який зростає, відповідно зростає і ймовірна вартість автомобіля. Для того щоб побачити можливу статистичну залежність можна побудувати парні графіки, що будує попарні залежності ознак набору даних.

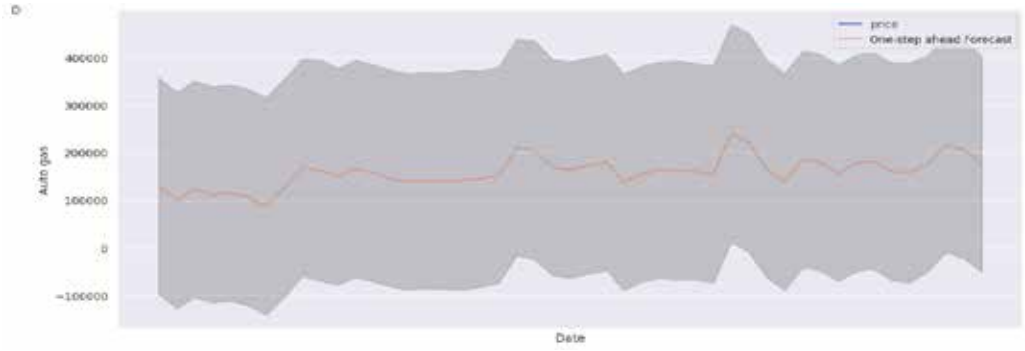

Рис. 8. Графік прогнозування

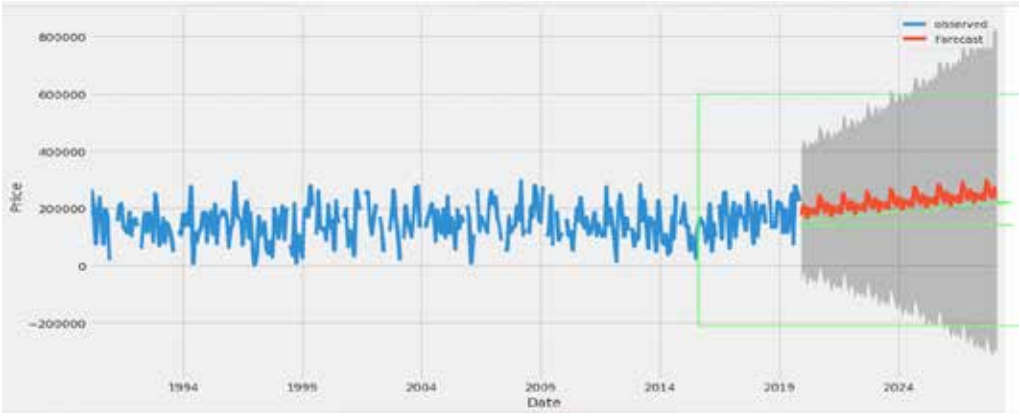

Рис. 9. Графік росту

Аналізуючи графік прогнозування (рис. 9) можна побачити коливання ціни в залежності від сезону та його ріст на майбутні роки, причому щорічно буде певний спад ціні в певному сезоні кварталі.

Проведене дослідження дозволяє зробити висновок про те, що обрана модель добре підходить для аналізу і прогнозування даних часових рядів.

Можна побачити періодичне зростання ціни і ці зростання повторюються з певним часом, що можна побачити на графіках тенденції та сезонності, а також на графіках тимчасових рядків. Можна з певною точністю сказати що взяті параметри є одними із ключових при формування ціни на продаж та купівлю авто.

Висновки. Для аналізу було сформовано джерело даних на 10 000 записів, за допомогою онлайн сервісу. Це дало змогу отримати більш точні результати порівняно з тим, якщо б записів було б значно менше, адже недостатність даних дає не повну картину. Була використана модель прогнозування авторегресійної інтегрованої ковзної середньої (ARIMA). Для вирішення поставленого завдання обрано алгоритм часових рядків. Мовою реалізації поставленого завдання було обрано Python. Отриманні результати дають можливість зробити висновки, що ціна автомобіля має сезонність зростання ціни, що дає можливість спрогнозувати свої дії та прийняти зважене рішення щодо купівлі або продажу авто.

Список літератури:

1. Levkivskyi, V., Lobanchykova, N., Marchuk, D. Research of algorithms of Data Mining. E3S Web of Conferences. Vol. 166. 05007. 2020. The International Conference on Sustainable Futures: Environmental, Technological, Social and Economic Matters (ICSF 2020). URL: https://doi.org/10.1051/e3sconf/202016605007.

2. Годлевський, Ю.О., Марчук, Г.В., Панаріна, І.В. 2022. Аналіз, моделювання та прогнозування ціни будинків залежно від їх розмірів. Технічна інженерія. 2(90) (Груд 2022), 79–86. DOI:https://doi.org/10.26642/ten-2022-2(90)-79-86.

3. Антонюк, Д.С., Вакалюк, Т.А., Марчук, Г.В., Дідківський, В.В. Прогнозування оцінки кредитоспроможності фізичних осіб із використанням можливостей ML. NET. Збірник наукових праць Національного університету кораблебудування імені Адмірала Макарова: Наукове видання. № 3 (481) 2020. – С. 63-71. DOI: https://doi.org/10.15589/znp2020.3(481).8.

4. Алгоритмічно-програмне забезпечення обробки та аналізу потоку кадрів відеоданих, що надходять з камер міста : комп'ютерна програма / В.Л. Левківський, Г.В. Марчук, В.В. Ципоренко, Д.К. Марчук. – 2021 [Electronic resource]. – Access mode : https://cutt.ly/jMxy1sq.

5. Surindar Gopalrao Wawale, Malik Jawarneh, P. Naveen Kumar, Thomas Felix, Jyoti Bhola, Roop Raj, Sathyapriya Eswaran, Rajasekhar Boddu, "Minimizing the Error Gap in Smart Framing by Forecasting Production and Demand Using ARIMA Model", Journal of Food Quality, vol. 2022, Article ID 1139440, 9 pages, 2022. https://doi.org/10.1155/2022/1139440

6. Chyon F. A., Suman M. N. H., Fahim M. R. I., Ahmmed M. S. , Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning, Journal of Virological Methods 301 (2022) 114433. URL: https://doi.org/10.1016/j.jviromet.2021.114433

7. Tandon H, Ranjan P, Chakraborty T, Suhag V. Coronavirus (COVID-19): ARIMA-based Time-series Analysis to Forecast near Future and the Effect of School Reopening in India. Journal of Health Management. 2022;24(3):373-388. doi:10.1177/09720634221109087

8. L. R. de Araújo Morais, G. S. da Silva Gomes, Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model, Applied Soft Computing 126 (2022) 109315. URL: https://doi.org/10.1016/j.asoc.2022.109315

9. Kiarie J., Mwalili S., Mbogo R., Forecasting the spread of the COVID-19 pandemic in Kenya using SEIR and ARIMA models, Infectious Disease Modelling 7 (2022) 179–188. URL: https://doi.org/10.1016/j.idm.2022.05.001

10. Zheng A, Fang Q, Zhu Y, Jiang C, Jin F, Wang X. An application of ARIMA model for predicting total health expenditure in China from 1978-2022. J Glob Health. 2020 Jun;10(1):010803. doi: 10.7189/jogh.10.010803. PMID: 32257167; PMCID: PMC7101215.

11. CSV generator, 2023. URL: https://extendsclass.com/csv-generator.html

Marchuk D.K., Kravchenko S.M., Levchenko A.Yu., Lezhnyov I.Ya. USING TIME SERIES IN FORECASTING THE CASH VALUE OF CARS

The article considers the problem of forecasting the monetary value of cars by parameters using time series. Currently, the task of forecasting is relevant for various practical applications. These can be such areas of activity as economy, finance, business, trade, politics, etc. Time series forecasting is decided on the basis of the created model that describes the process under study. The object of the study is the application of autoregressive integrated moving average (ARIMA) models for forecasting the value of a car using time series. ARIMA models are capable of simulating a wide range of seasonal data. In the process of research, an own dataset was created with the help of an online service, which consisted of 10,000 records, the data of which was saved in CSV format. The Python programming language was chosen as the implementation tool, namely the following libraries: pandas, numpy, matplotlib, seaborn, statsmodels, itertools and keras. Research results are presented visually. The analysis was carried out according to various parameters: the cost of the car, the year of manufacture, the type of gearbox, and the type of fuel. The analysis was carried out using various ARIMA tools and methods. As a result of the analysis, it can be concluded that the selected model satisfies the conditions of time series forecasting. Analyzing the graphs, you can see the seasonality and rapid growth in the cost of cars for the coming years, and every year there will be a certain decline in the price in a certain season of the quarter. Based on the results of seasonality and trends, it can be concluded that the price of a car has a seasonal increase in price, which makes it possible to predict your actions and make an informed decision about buying or selling a car.

Key words: *car, forecasting, time series, ARIMA, trend, seasonality.*