УДК 004.8:330.3:339.5 https://doi.org/10.35546/kntu2078-4481.2022.4.7

С. М. КОНЦЕБА
Уманський національний університет садівництва
ORCID: 0000-0003-4161-5581
Р. І. ЛІЩУК
Уманський національний університет садівництва
ORCID: 0000-0002-2051-5365

ВИКОРИСТАННЯ МЕТОДІВ DATA MINING ДЛЯ ПРОГНОЗУВАННЯ ПОКАЗНИКІВ ЗОВНІШНЬОЕКОНОМІЧНОЇ ДІЯЛЬНОСТІ

У статті описані результати дослідження використання алгоритмів машинного навчання для аналізу і прогнозування показників зовнішньоекономічних операцій в Україні. Метою цієї статті є прогнозування показників імпорту і експорту з використанням алгоритмів машинного навчання (лінійна регресія, Gaussian Process Regression, SMOreg і нейронна мережа Multilayer Perceptron) на статистичних даних, що охоплюють період з 1 січня 2018 р. по 31 грудня 2021 р. З метою виявлення найточнішого результату прогнози зроблені з використанням статистичних даних для різних інтервалів базового періоду та періодів прогнозування. Точність алгоритмів машинного навчання оцінювалася за допомогою порівняння наступних показників: середня абсолютна похибка (MAE), середньоквадратична похибка (RMSE), та середня абсолютна похибка у відсотках (MAPE). Розраховані прогнозні показники зовнішньоекономічних операцій за алгоритмом SMOreg мають високу точність прогнозу, оскільки мають найменші показники абсолютної похибки у відсотках (MAPE). Показники середньої абсолютної похибки (MAE) і середньоквадратичної похибки (RMSE) також вказують що алгоритм SMOreg має високу точність прогнозу. Результати аналізу показали, що алгоритми машинного навчання досягли високоточної ефективності прогнозування. Виявлено, що нелінійні моделі значно краще справляються із задачею прогнозування експортно-імпортних операцій, ніж лінійні моделі. Загальна точність алгоритму SMOreg була кращою для всього інтервалу базового періоду та вибраного періоду прогнозу. Результати, отримані в результаті цього аналізу, можуть допомогти фахівцям з економіки в оцінці показників зовнішньоекономічних операцій в Україні. Реалізація прогнозування експортно-імпортних операцій на підставі використання алгоритму SMOreg може бути автоматизована для створення експертної системи з метою оцінки показників зовнішньоекономічних операцій в розрізі окремих регіонів.

Ключові слова: *Data Mining, лінійна регресія, Gaussian Process Regression, SMOreg алгорит, Multilayer Perceptron, прогнозування, зовнішньоекономічна діяльність.*

S. M. KONTSEBA
Uman National University of Horticulture
ORCID: 0000-0003-4161-5581
R. I. LISHCHUK
Uman National University of Horticulture
ORCID: 0000-0002-2051-5365

USAGE OF DATA MINING METHODS FOR FORECASTING FOREIGN ECONOMIC ACTIVITY INDICATORS

The article describes the results of the study of machine learning algorithms usage for foreign economic transactions in Ukraine analysis and forecasting. The purpose of this article is to forecast import and export indicators using machine learning algorithms (linear regression, Gaussian Process Regression, SMOreg and Multilayer Perceptron neural network) based on statistics covering the period from January 1, 2018 to December 31, 2021. The most accurate forecast result was identified by using statistics for different intervals of the base period and forecast periods. The accuracy of machine learning algorithms was assessed by comparing the following indicators: mean absolute error (MAE), root mean square error (RMSE), and mean absolute error in percent (MAPE). The calculated forecast indicators of foreign economic operations according to the SMOreg algorithm have high forecast accuracy due to the smallest indicators of the mean absolute percentage error (MAPE). The mean absolute error (MAE) and the root mean square error (RMSE) also indicate that the SMOreg algorithm has high prediction accuracy. The results of the analysis showed that machine learning algorithms have achieved highly accurate forecasting efficiency. It was found that nonlinear models cope much better with forecasting export-import operations than linear models. The overall accuracy of the SMOreg algorithm was better for the entire base period interval and the selected forecast period. The results of this analysis can help economic experts in assessing the performance of foreign economic transactions in Ukraine. Implementation of forecasting of export-import operations based on the use of the SMOreg algorithm can be automated to create an expert system to assess the performance of foreign economic transactions in terms of individual regions.

Key words: *Data Mining, linear regression, Gaussian Process Regression, SMOreg, Multilayer Perceptron, forecasting, foreign economic activity.*

Постановка проблеми

Фахівцям з економіки в Україні необхідно динамічно і швидко оцінювати показники зовнішньоекономічних операцій особливо в період нестабільної економічної ситуації. Тому для організацій та установ постає потреба мати адекватну модель прогнозування експортно-імпортних операцій. Подібні лінійні моделі вже використовуються організаціями та установами, які працюють у сфері зовнішньоекономічної діяльності. Однак як показує практика нелінійні моделі значно краще справляються із задачею прогнозування експортно-імпортних операцій, ніж лінійні моделі. Тому використання методів Data Mining дасть можливість створити адекватну модель прогнозування показників зовнішньоекономічної діяльності.

Аналіз останніх досліджень і публікацій

Проблеми прийняття рішень в економічних системах на основі прогнозування показників з використанням методів інтелектуального аналізу даних розглянуто у працях вітчизняних авторів Р.О. Петрова, О.Я. Кучерука [1]. Прогнозуючи терміни продажу товарів, звертають увагу на те, що для прогнозування продажів найчастіше використовуються класичні методи аналізу. часових рядів та дерева рішень. К.В. Ілляшенко [1] використовує методи Data Mining для аналізу великих обсягів показників у бухгалтерському обліку. П.І. Бідюк, С.М. Савченко, А.С. Савченко [3] визначили переваги методів інтелектуального аналізу та їх комбінацій (гібридні методи інтелектуального аналізу даних) для прийняття раціональних рішень у системі управління та прогнозування конкурентоспроможності вітчизняних підприємств. Група вчених Л.О. Коршевнюк [4], Г. Чорноус, С. Рибальченко [5], О.Ю. Берзлев, М.М. Маляр, В.В. Ніколенко [6], П.І. Бідюк, А.В. Федоров [7], О.М. Міхайлуца, А.В. Пожуєв, В.В. Тищенко [8], застосовують методи інтелектуального аналізу даних для прогнозування біржових показників, процесів ціноутворення, оцінювання фінансових ризиків та електронної комерції. Група іноземних вчених Б. Жмук, Х. Йошич [9] та Д. Асір Ентоні Гнана Сінгх, Е. Джебамалар Лівлайн, С. Мутукрішнан, Р. Юварадж [10] на основі алгоритмів машинного навчання прогнозували індекси фондового ринку та бізнес-показники міжнародних компаній.

Метою статті є прогнозування показників імпорту і експорту з використанням алгоритмів машинного навчання (лінійна регресія, Gaussian Process Regression, SMOreg і нейронна мережа Multilayer Perceptron) на статистичних даних.

Викладення основного матеріалу дослідження

Лінійна регресія (LReg) – це найпростіша модель для прогнозування, що зв'язує залежну змінну $|y| = p$ з помилкою $|\varepsilon| = p$ з незалежними змінним $|x| = n$ x p, яку можна записати як:

$$y = {}^{2}_{0} + \sum_{i=1}^{p} {}^{2}_{j} x_j + \mu$$

де y – теоретичне значення вектору-стовпчика результативної ознаки, який має розмінність n $(n > p)$

x_j – аргументи (фактори);

n – число досліджуваних факторів;

β_j- коефіцієнти регресії, що показують ступінь впливу кожного з факторів на функцію;

β_0 – залишковий член, що характеризує середнє значення функції.

Gaussian Process Regression (GPreg) докладно описана Расмуссеном і Вільямсом [11, с. 13]. Вони використали метод Гауса (GP) для опису розподілу за функціями. Формально метод Гауса – це сукупність випадкових величин, кінцеве число яких має спільний багатовимірний нормальний розподіл.

Метод Гауса повністю специфікований за своєю середньою функцією $m(x)$ та функцією коваріації $k(x; x')$ для реального процесу $f(x)$ як

$$m(x) = E\left[f(x)\right],$$

$$k(x, x') = E\left[f(x) - m(x))\left(f(x') - m(x')\right)\right],$$

і використавши метод Гауса може бути записаний:

$$f(x) \sim GP\left(m(x), k(x, x')\right).$$

SMOreg алгоритм. SMOreg розшифровується як послідовна мінімальна оптимізація. Це алгоритм реалізації методу опорних векторів (SVM) для регресії. В основному він використовується для навчання SVM. Навчання SVM відбувається шляхом вирішення дуже великої задачі оптимізації квадратичного програмування:

$$\max_{a} \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K\left(x_i, x_j\right) a_i a_j$$

при умові $0 \le a_i \le C$, для i=1,2,…,n

$$\sum_{i=1}^{n} y_i a_i = 0$$

де C – гіперпараметр SVM;

$K(x_i, x_j)$ – функція ядра;

змінні a_i є множниками Лагранжа [12, с. 13].

SMO ірраціональний алгоритм вирішення такої задачі. SMO спочатку розділяє велику проблему QP на множину невеликих підзадач QPproblem, які потім розв'язуються аналітично:

Через обмеження лінійної рівності, що включає множники Лагранжа a_i, найменша можлива задача включає два такі множники. Тоді для будь-яких двох множників, обмеження зводяться до:

$$0 \le a_1 a_2 \le C,$$

$$y_1 a_1 + y_2 a_2 = k$$

Ця задача може бути вирішена аналітично: потрібно знайти мінімум одновимірної квадратичної функції k є негативним значенням суми за іншими умовами в обмеженні рівності.

Multilayer Perceptron (MLP). Мережа MLP характеризується декількома шарами вхідних вузлів, з'єднаних між собою прямим шляхом. Мережі MLP – це штучні нейронні мережі, які для побудови вихідного блоку використовують просту модель персептрона. Топологія складається з шарів паралельних персептронів з оптимальними зв'язками між шарами:

$$y_t = \beta_0 + \sum_{j=1}^{q} \beta_j g\left(\sum_{i=1}^{p} w_{ij} x_{t-1} + w_{0j}\right)$$

$$\{x_{t-1}; i = 1, \ldots, p\} \quad (1)$$

$$\{w_{ij}; i = 1, \ldots, p; j = 1, \ldots, q\}$$

$$\{\beta_j; i = 1, \ldots, q\}$$

де y_t – вихідний вектор мережі в момент часу t;
x_{t-i} – вхідне значення в момент часу t-i;
β_j – вага з'єднання виходу нейрона j на прихованому шарі з вихідним нейроном;.
w_{ij} – вага з'єднання нейрона j з входом прихованого шару;
і g – нелінійна функція нейронів у прихованому шарі.

Кількість нейронів у прихованому шарі позначається q і визначає мережеву здатність апроксимувати задану функцію [13, с. 11].

Для оцінки точності моделей у програмі WEKA за замовчуванням її обчислюються середня абсолютна похибка (MAE) і середньоквадратична похибка (RMSE), але для більш точного обчислення ефективності алгоритму навчання в наборі даних було додатково вибрано обчислення показників середньої абсолютної похибки у відсотках (MAPE).

Визначення показника середньої абсолютної похибки (MAE) використовуються для оцінки результату. Він показує найближчі значення величини на основі прогнозу кінцевого результату. Середня абсолютна помилка (MAE) – це величина, яка використовується для вимірювання того, наскільки прогнози близькі до кінцевих результатів. Визначають показник за наступною формулою:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} x_i - x$$

де n – кількість похибок;
$x_i - x$ – абсолютні похибки.

Стандартні статистичні показники RMSE використовуються для вимірювання ефективності алгоритмів щодо вибраного для дослідження набору даних. Він порівнює передбачене значення та відоме значення і визначається за формулою:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{F_i - A_i}{n}}$$

де F_i – значення прогнозу;
A_i – фактичне значення.

Середня абсолютна похибка у відсотках (MAPE) – це статистичний показник, що є мірою точності прогнозу. Це найпоширеніший показник, що використовується для оцінки точності прогнозування або як функція втрат для проблем регресії в машинному навчанні. Зазвичай виражається як відношення, яке визначається за формулою:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{A_i - F_i}{A_i}$$

Наведені вище показники виражають коефіцієнти похибок прогнозування у відсотках. Чим меншими є значення показників, тим більша точність прогнозу.

Набір даних становить показники зовнішньоекономічних операцій в Україні, а саме експорту і імпорту, за період 2018–2021 роки в розрізі місяців [54]. Таким чином розмір вибірки становив 48 одиниць. Набір даних представлено у форматі файлу .arff, імпортується до WEKA версії 3.9.5, мітка часу встановлюється

як «Key_Data», періодичність як «Квартал», кількість одиниць часу для прогнозування дорівнює 4, а прапорець «Виконати оцінку» встановлено з метою оцінити рівень точності алгоритму. Після цього було здійснено тестування чотирьох алгоритмів, щоб знайти той, який найкраще описує набір даних, використовуючи три вибрані показники точності (табл. 1).

Таблиця 1

Значення показників оцінки точності прогнозу зовнішньоекономічних операцій в Україні, %

	Експорт			Імпорт		
Період прогнозування	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**
Gaussian Process Regression						
I кв. 2022	337,6	8,1	445,1	468,4	9,8	608,8
II кв. 2022	410,6	10,0	562,6	537,6	11,4	688,7
III кв. 2022	462,0	11,2	634,1	566,9	12,0	734,1
IV кв. 2022	529,4	12,7	721,6	608,1	13,0	794,0
Лінійна регресія						
I кв. 2022	234,9	5,5	313,1	375,7	7,7	480,4
II кв. 2022	374,5	8,4	485,6	474,4	9,6	586,3
III кв. 2022	545,7	11,9	681,4	557,3	11,1	682,8
IV кв. 2022	766,1	16,2	948,0	691,9	13,5	823,7
Multilayer Perceptron (MLP)						
I кв. 2022	284,2	6,0	387,2	366,8	7,2	449,8
II кв. 2022	448,7	9,5	642,1	635,5	12,1	790,4
III кв. 2022	538,9	11,7	743,5	856,3	16,4	1051,6
IV кв. 2022	739,2	16,0	1025,2	1050,9	20,0	1293,2
SMOreg алгоритм						
I кв. 2022	226,3	5,6	318,0	341,3	7,2	478,4
II кв. 2022	331,9	7,9	435,3	404,5	8,6	538,4
III кв. 2022	414,8	9,6	539,3	435,5	9,2	559,2
IV кв. 2022	567,4	12,7	688,2	477,4	10,1	594,6

Значення середньої абсолютної похибки у відсотках (MAPE) для прогнозованих значень експортних операцій за алгоритмом Gaussian Process Regression знаходиться в межах 8,1 – 12,7%, за методом лінійної регресії в межах 5,5 – 16,2%, за алгоритмом Multilayer Perceptron в межах 6,0 – 16,0%, за алгоритмом SMOreg в межах 5,6 – 12,7%.

Значення середньої абсолютної похибки у відсотках (MAPE) для прогнозованих значень імпортних операцій за алгоритмом Gaussian Process Regression знаходиться в межах 9,8 – 13,0%, за методом лінійної регресії в межах 7,7 – 13,5%, за алгоритмом Multilayer Perceptron в межах 7,2 – 20,0%, за алгоритмом SMOreg в межах 7,2 – 10,1%.

Таким чином розраховані прогнозні показники зовнішньоекономічних операцій за алгоритмом SMOreg мають високу точність прогнозу, оскільки мають найменші показники абсолютної похибки у відсотках (MAPE).

Аналогічні висновки можна зробити проаналізувавши показники середньої абсолютної похибки (MAE) і середньоквадратичної похибки (RMSE).

Отже, алгоритму SMOreg є найбільш прийнятним для прогнозування зовнішньоекономічних операцій. Загальна точність алгоритму SMOreg була кращою для всього інтервалу базового періоду та вибраного періоду прогнозу.

Також показників точності використаних алгоритмів вказує на те, що нелінійні моделі значно краще справляються із задачею прогнозування експортно-імпортних операцій, ніж лінійні моделі.

Результати обробки даних зовнішньоекономічної діяльності підприємств і організацій України за період, що досліджувався, за допомогою алгоритму SMOreg та поквартальні прогнозні показники на 2022 рік наведено на рис. 1.

Рис. 1. Результати обробки даних зовнішньоекономічної діяльності підприємств і організацій України за допомогою алгоритму SMOreg та поквартальні прогнозні показники на 2022 рік

Висновки

Виявлено, що нелінійні моделі значно краще справляються із задачею прогнозування експортно-імпортних операцій, ніж лінійні моделі. Загальна точність алгоритму SMOreg була кращою для всього інтервалу базового періоду та вибраного періоду прогнозу. В результаті побудованого прогнозу показників експорту отримали наступні результати: I квартал 2022 року 7006,5 млн. грн., II квартал 2022 року 7389,3 млн. грн., III квартал 2022 року 7907,3 млн. грн., IV квартал 8411,5 млн. грн. Прогнозні показники імпорту наступні: I квартал 2022 року 7799,7 млн. грн., II квартал 2022 року 8133,0 млн. грн., III квартал 2022 року 8296,6 млн. грн., IV квартал 8357,3 млн. грн. Результати, отримані в результаті цього аналізу, можуть допомогти фахівцям з економіки в оцінці показників зовнішньоекономічних операцій в Україні. Реалізація прогнозування експортно-імпортних операцій на підставі використання алгоритму SMOreg може бути автоматизована для створення експертної системи з метою оцінки показників зовнішньоекономічних операцій в розрізі окремих регіонів.

Список використаної літератури

1. Петров Р.О., Кучерук О.Я., Прогнозування термінів продажу товарів методами інтелектуального аналізу даних. *Актуальні проблеми компютерних наук*. 2019. URL: http://elar.khnu.km.ua/jspui/bitstream/123456789/7933/1/APKN-2019_%28v_2_0%29-156-158.pdf (дата звернення 07.12.2022).

2. Ілляшенко К.В. Використання методів DATA MINING у бухгалтерському обліку. *Бухгалтерський облік, аналіз та аудит*. 2019. Випуск 6(17). С. 347–376.

3. Бідюк П.І., Савченко С.М., Савченко А.С., Методи інтелектуального аналізу даних в прогнозуванні конкурентоспроможності підприємств. 2018. URL: http://www.ei-journal.in.ua/index.php/journal/article/view/61/48 (дата звернення 20.01.2021).

4. Коршевнюк Л.О., Бідюк П.І., Інформаційно-аналітична система для адаптивного прогнозування фінансових процесів та оцінювання ризиків. *Наукові праці. Комп'ютерні технології*. 2013. Вип. 201, т. 213, С. 59–62.

5. Чорноус Г., Рибальченко С.. Оптимізація ціноутворення на основі моделей інтелектуального аналізу даних. *Вісник Київського національного університету імені Тараса Шевченка*. 2015. № 7 (172), С. 52–58.

6. Берзлев О.Ю, Маляр М.М., Ніколенко В.В. Адаптивні комбіновані моделі прогнозування біржових показників. *Вісник Черкаського держ. технолог. унту. Серія: технічні науки*. 2011. № 1. С. 50–54.

7. Бідюк П.І., Федоров А.В.. Ймовірнісне прогнозування процесів ціноутворення на фондових ринках. *Системні дослідження та інформаційні технології*. 2009. № 1. С. 65–73.

8. Міхайлуца О.М., Пожуєв А.В., Тищенко В.В. Методи інтелектуального аналізу даних та їх застосування в електронній комерції. *Математичне моделювання*. 2020. № 1(42). С. 154–163.

9. Berislav Žmuk, Hrvoje Jošić. Forecasting stock market indices using machine learning algorithms. *Interdisciplinary Description of Complex Systems*, 2020. №18(4). P. 471–489.

10. D. Asir Antony Gnana Singh, E. Jebamalar Leavline, S. Muthukrishnan, R. Yuvaraj. Machine Learning based Business Forecasting. *I.J. Information Engineering and Electronic Business*, 2018, № 6, p. 40–51.

11. Rasmussen C. E., Williams C. K. I. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005.

12. Шеремет О. І., Садовой О. В. Метод опорних векторів (SVM). Математичне моделювання. 2013. № 1(28). С. 13–17.

13. Oscar Claveria, Enric Monte, Salvador Torra. Regional tourism demand forecasting with machine learning models: Gaussian process regression vs. neural network models in a multiple-input multiple-output setting. Barcelona: Institut de Recerca en Economia Aplicada Regional i Pública, 2017. 23 c.

References

1. Petrov R.O., Kucheruk O.Y. (2019) Prohnozuvannia terminiv prodazhu tovariv metodamy intelektualnoho analizu danykh. Aktualni problemy kompiuternykh nauk. URL: http://elar.khnu.km.ua/jspui/bitstream/123456789/7933/1/APKN-2019_%28v_2_0%29-156-158.pdf (data zvernennia 07.12.2022).

2. Illiashenko K.V. (2019) Vykorystannia metodiv DATA MINING u bukhhalterskomu obliku. *Bukhhalterskyi oblik, analiz ta audit,* vol. 6(17), pp. 347–376.

3. Bidiuk P.I., Savchenko S.M., Savchenko A.S. (2018) Metody intelektualnoho analizu danykh v prohnozuvanni konkurentospromozhnosti pidpryiemstv. URL: http://www.ei-journal.in.ua/index.php/journal/article/view/61/48 (data zvernennia 20.01.2021).

4. Korshevniuk L.O., Bidiuk P.I. (2013) Informatsiino-analitychna systema dlia adaptyvnoho prohnozuvannia finansovykh protsesiv ta otsiniuvannia ryzykiv. *Naukovi pratsi. Komp'iuterni tekhnolohii,* vol. 201, t. 213, pp. 59–62.

5. Chornous H., Rybalchenko S. (2015) Optymizatsiia tsinoutvorennia na osnovi modelei intelektualnoho analizu danykh. *Visnyk Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka,* no. 7 (172), pp. 52–58.

6. Berzlev O.Y., Maliar M.M., Nikolenko V.V. (201) 1Adaptyvni kombinovani modeli prohnozuvannia birzhovykh pokaznykiv. *Visnyk Cherkaskoho derzh. tekhnoloh. untu. Seriia: tekhnichni nauky,* no. 1, pp. 50–54.

7. Bidiuk P.I., Fedorov A.V. (2009) Ymovirnisne prohnozuvannia protsesiv tsinoutvorennia na fondovykh rynkakh. *Systemni doslidzhennia ta informatsiini tekhnolohii*, no. 1, pp. 65–73.

8. Mikhailutsa O.M., Pozhuiev A.V., Tyshchenko V.V. (2020) Metody intelektualnoho analizu danykh ta yikh zastosuvannia v elektronnii komertsii. *Matematychne modeliuvannia*, no 1(42), pp. 154–163.

9. Berislav Žmuk, Hrvoje Jošić. Forecasting stock market indices using machine learning algorithms. *Interdisciplinary Description of Complex Systems*, 2020. №18(4). P. 471–489.

10. D. Asir Antony Gnana Singh, E. Jebamalar Leavline, S. Muthukrishnan, R. Yuvaraj. Machine Learning based Business Forecasting. *I.J. Information Engineering and Electronic Business*, 2018, № 6, p. 40–51.

11. Rasmussen C. E., Williams C. K. I. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005.

12. Sheremet O. I., Sadovoi O. V. (2013) Metod opornykh vektoriv (SVM). *Matematychne modeliuvannia*, no. (28), pp. 13–17.

13. Oscar Claveria, Enric Monte, Salvador Torra. Regional tourism demand forecasting with machine learning models: Gaussian process regression vs. neural network models in a multiple-input multiple-output setting. Barcelona: Institut de Recerca en Economia Aplicada Regional i Pública, 2017. 23 c.