УДК 004.85, 004.89, 004.9

ДОСЛІДЖЕННЯ ТА ЕКСПЕРИМЕНТАЛЬНИЙ АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ В ЗАДАЧАХ ЕЛЕКТРОННОЇ КОМЕРЦІЇ

П. Б. Вітинський[1], Р. О. Ткаченко[1], Б. М. Гавриш[2]

[1]Національний університет «Львівська політехніка», вул. С. Бандери, 12, Львів, 79013, Україна

[2]Українська академія друкарства, вул. Під Голоском, 19, Львів, 79020, Україна

Розглянуто задачу аналітики великих даних у персоналізованих системах електронної комерції. Здійснено огляд наявних методів машинного навчання для розв'язання задач регресії в таких системах. Окреслено переваги та недоліки розглянутих методів та алгоритмів. Проведено моделювання роботи методів машинного навчання для прогнозування суми витрат споживачів роздрібного магазину. Експериментально визначено середньоквадратичну похибку результатів прогнозування. Встановлено, що найвищу точність роботи забезпечують методи на основі ансамблювання — Random forest та AdaBoost. Проведено оцінювання тривалості процедур навчання усіх досліджуваних методів. Встановлено, що найбільша швидкість забезпечується під час використання методу стохастичного градієнтного спуску. Виявлено, що ефективність застосування нейромережевих методів не є задовільною як з огляду на точність, так і на час їх роботи.

Ключові слова: *ансамбль, електронна комерція, машинне навчання, прогнозування, random forest, задача регресії, штучний інтелект.*

Постановка проблеми. Розвиток комерційної діяльності засобами електронно-інформаційного бізнесу забезпечує чимало переваг як для споживачів подібних послуг, так і для компаній, які створюють цю пропозицію [1]. Освоєння нових ринків товарів та послуг інтернет-компаніями забезпечує інтернет-споживачу великі можливості для вивчення та купівлі нової продукції. Бурхливий розвиток комп'ютерної техніки та інформаційних систем дає змогу користувачу купувати продукцію навіть із телефона. Проте частка роздрібних продажів із використанням подібних систем все ще є дуже малою. Традиційна модель електронної комерції, що ґрунтується на пошуку товару чи послуги в Інтернеті стає вузьким майданчиком, що перешкоджає подальшому розвитку компанії. На зміну їй приходить нова, персоналізована система, що забезпечує низку переваг [2].

Розвиток сучасних інтелектуальних систем електронної комерції потребує точного та швидкого розв'язку задач прогнозування. Ефективний інтелектуальний аналіз історичних даних про купівлю товарів і послуг різними групами споживачів, що ґрунтується на персоналізованому підході, надає низку переваг для

інтернет-компаній. Серед очевидних — збільшення рівня продажів на основі вчасного та правильного визначення потенційних потреб споживача і, як наслідок, збільшення прибутку. Це стає можливим із використанням ефективної моделі побудови прогнозів на основі низки характеристик того чи іншого споживача [3].

Аналіз останніх досліджень та публікацій. Задачі побудови нових та розвитку наявних систем електронної комерції, що ґрунтуються на поєднанні персоналізованого підходу та методів штучного інтелекту набули значного розвитку. У публікації [4] аналізується застосування інструментарію Data Mining для розв'язання задач електронної комерції, що ґрунтується на традиційній моделі. Окрім цього, автор розглядає лише деякі алгоритми прогнозування, що в кінцевому підсумку дають змогу збільшити прибутки фірми. Автори у праці [2] розглянули персоналізований підхід до побудови систем електронної комерції. Дослідження цілої групи інших авторів у цьому напрямі підтверджують його ефективність [5]. Проте в усіх вищевказаних працях недостатньо уваги приділено дослідженню та аналізу різноманітних методів і алгоритмів машинного навчання для розв'язання актуальної задачі роздрібного продажу на основі персоналізованого підходу в системах електронної комерції. Результат такої роботи забезпечить точне та вчасне визначення потреб як наявних, так і потенційних споживачів [6], що значно збільшить якість сервісу з одного боку та прибутки компанії — з іншого.

Мета статті — дослідження та експериментальний аналіз застосування сучасних методів машинного навчання для розв'язання задач регресії в галузі електронної комерції.

Дослідження методів машинного навчання для розв'язання задач електронної комерції. Методи машинного навчання для розв'язання задачі регресії в галузі електронної комерції, які досліджувались у роботі, подано у табл. 1.

Перша група методів — ансамблеві. У публікації висвітлено дослідження двох різних класів цих методів: bagging, до якого належить алгоритм Random forest, та boosting, представником якого є алгоритм AdaBoost. Особливістю Random forest є те, що він забезпечує стійке та ефективне рішення при мінімізації проблем із перенавчанням [7]. Окрім цього, він стійкий до викидів і масштабування та здатен опрацьовувати чималі набори даних із великою кількістю ознак кожного вхідного вектора [8]. Серед недоліків цього алгоритму потрібно зазначити відсутність екстраполятивних властивостей та складну інтерпретацію результатів [7].

Композиція алгоритму AdaBoost передбачає ітеративний процес побудови часткових моделей, де кожен наступний алгоритм навчається із використанням інформації про похибки, зроблені на попередньому етапі [9]. Простота реалізації та висока здатність до генералізації — це основні переваги алгоритму AdaBoost. Серед обмежень алгоритму потрібно зазначити необхідність відсутності шумів у даних, які можуть призвести до перенавчання [10]. Окрім цього, важливу роль щодо ефективності роботи цього алгоритму відіграє розмірність вибірки для навчання [9].

Наступна група методів розв'язання задачі регресії, яка досліджувалася, — нейромережева. Розглянуто можливість застосування багатошарового перцептрону, нейронної мережі узагальненої регресії та нейроподібної структури моделі

послідовних геометричних перетворень для прогнозування суми витрат споживачів роздрібного магазину. Незважаючи на можливість достатньо точної апроксимації, багатошаровий перцептрон характеризується довготривалістю процедури навчання через її ітеративний процес. Нейронна мережа узагальненої регресії швидка, проте такі характеристики, як розмір і структура вибірки даних, якість алгоритму та програмне рішення на його основі, застосування паралелізму тощо, в деяких випадках роблять мережу дуже великою і повільною. Окрім цього, як і алгоритм Random forest, вона не здатна до екстраполяції даних [11]. Застосування нейроподібних структур моделі послідовних геометричних перетворень до розв'язання задач регресії характеризується високою швидкістю реалізації процедур навчання та достатньою точністю прогнозу [12]. Проте великі обсяги даних можуть накладати обмеження на застосування цього обчислювального інструменту.

Таблиця 1

Параметри досліджуваних методів машинного навчання

№ з/п	Назва методу машинного навчання	Умовне позначення методу	Параметри методу
1	Random Forest	Метод 1	максимальна глибина кожного дерева = 5
2	Алгоритм AdaBoost	Метод 2	базовий алгоритм — дерево рішень (максимальна глибина = 4), кількість слабких (базових) дерев = 300
3	Багатошаровий перцептрон	Метод 3	23 входи, 23 нейрони в прихованому шарі, 1 вихід
4	Нейроподібна структура МПГП	Метод 4	23 входи, 23 нейрони в прихованому шарі, 1 вихід
5	Лінійна регресія на основі стохастичного градієнтного спуску	Метод 5	функція втрат = 'squared_loss', α=0.0001
6	Нейронна мережа узагальненої регресії	Метод 6	$\sigma = 0.4$ ($\sigma \in [0.1,1.5]$
7	Регресор на основі Машини опорних векторів	Метод 7	ядро = rbf, epsilon = 0.001, максимальна кількість ітерацій = 200

Лінійна регресія на основі стохастичного градієнтного спуску, так як і машина опорних векторів, характеризується високою швидкістю роботи [13], проте не завжди задовільними результатами прогнозування [14].

Експериментальний аналіз застосування методів машинного навчання для розв'язання задач електронної комерції. Моделювання роботи наявних методів машинного навчання (табл. 1) відбувалося на реальному наборі даних задачі електронної комерції [15]. Задача полягає у прогнозуванні суми витрат покупців у «Чорну п'ятницю» на основі низки ознак. Дані отримано з магазину роздрібної торгівлі. Детальний статистичний аналіз заданої вибірки подано у [16].

Моделювання відбувалося на наборі даних, із якого було вилучено усі спостереження із пропусками. Отже, вибірку даних (обсягом 15 673 вектори) було випадково розділено на навчальну та тестову. Відсоткове співвідношення такого поділу становило 70 % до 30 % відповідно. Навчальна вибірка складалася із 10 971 спостереження, кожне з яких містило 23 незалежні змінні та одну вихідну — суму покупки. Розмірність тестової вибірки становила відповідно 4 702 спостереження.

Усі параметри досліджуваних методів подано в табл. 1. Оцінка ефективності задачі прогнозування відбувалася на основі двох показників:

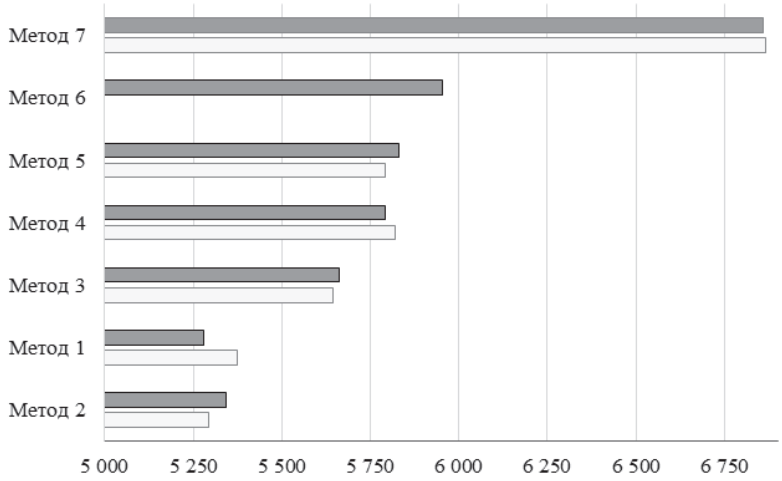
1. Середньоквадратична похибка (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_i^{pred} - y_i^{true})^2}. \quad (1)$$

2. Тривалість процедури навчання.

На рис. 1 подано порівняння значень середньоквадратичної похибки усіх досліджуваних методів.

На осі *ox* покладено значення похибки, на осі *oy* — досліджувані методи. Зеленими (темними) стовпцями гістограми позначено похибку, отриману в режимі навчання (RMSE навч.), жовтими (світлими), відповідно, похибку режиму застосування (RMSE тест).



	Метод 2	Метод 1	Метод 3	Метод 4	Метод 5	Метод 6	Метод 7
RMSE тест.	5343,21	5280,83	5660,74	5792,91	5830,01	5954,60	6858,63
RMSE навч.	5293,87	5375,84	5646,11	5821,64	5791,16		6865,32

Рис. 1. Величини похибок (RMSE) режимів навчання і застосування усіх досліджуваних методів машинного навчання (табл. 1)

Найгірший результат щодо точності розв'язання поставленої задачі демонструє регресор, побудований на основі машини опорних векторів (рис. 1). Найкращий результат отримано під час використання алгоритму AdaBoost. Різниця у

точності на основі (1) між обома методами становить більш ніж 28 %. Алгоритм Random forest порівняно із AdaBoost демонструє дещо гірші результати в режимі застосування, проте значно кращі в режимі навчання. Необхідно зазначити, що на рис. 1 не наведено похибку в режимі навчання для методу 6, оскільки навчання як такого він не потребує. Проте, як видно з рис. 1, точність його роботи є незадовільною.

Окрім точності роботи, не менш важливою характеристикою систем електронної комерції є можливість працювати в online-режимі. Саме тому тривалість процедур навчання методів, закладених в основі подібних систем, є критично важливою. З огляду на це, у публікації проведено експериментальне дослідження щодо оцінки цього показника.

На рис. 2 наведено тривалості процедур навчання для усіх досліджуваних методів (в секундах).

Як видно з рис. 2, процедура навчання *методу 5* є дуже швидкою, а тривалість алгоритмів навчання, закладених в основі *методів 1 та 2*, відповідно, в 6 та 12 разів повільніша. Застосування багатошарового перцептрона для розв'язання поставленої задачі призводить до чималих часових затримок. Зокрема, він повільніший за метод 1 більш ніж у 549 разів. Проте найгірші результати, з огляду на час роботи, отримано із використанням нейронної мережі узагальненої регресії. Її застосування для розв'язання поставленої задачі триває понад 130 секунд.

На основі аналізу як точності роботи методів машинного навчання (рис. 1), так і тривалості процедур їх навчання (рис. 2) можна стверджувати, що найефективніший розв'язок поставленої задачі забезпечують методи ансамблювання за схемами як покращеного об'єднання (bagging), так і покращеного перетину (boosting). Тобто *методи 1 та 2* відповідно. Штучні нейронні мережі (*методи 3 та 4*) тут не забезпечують достатньої точності.

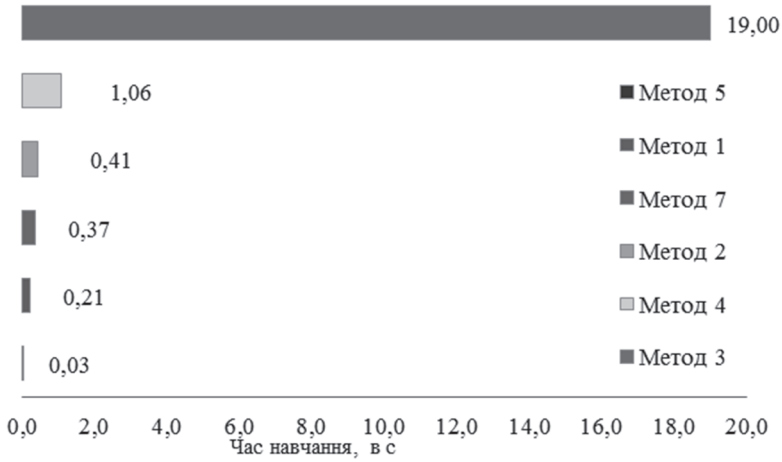


Рис. 2. Тривалість процедур навчання досліджуваних методів (табл. 1)

Подальша робота буде проводитися в напрямі ансамблювання нейронних мереж, зокрема, побудованих на основі моделі послідовних геометричних перетворень

для підвищення точності розв'язання задач регресії та класифікації в галузі електронної комерції.

Висновки. Досліджено основні переваги та недоліки застосування наявних методів машинного навчання для вирішення завдань електронної комерції. Проведено експериментальний аналіз їх застосування для розв'язання задачі прогнозування суми витрат споживачів роздрібного магазину на реальному наборі даних. Експериментально визначено точність роботи та час процедур навчання усіх досліджуваних методів. Незважаючи на високі часові характеристики роботи, методи на основі машини опорних векторів та стохастичного градієнтного спуску не забезпечують достатньої точності роботи. Ефективність застосування нейромережевих методів також не є задовільною як з огляду на точність, так і на час їх роботи. Встановлено, що найвищу точність розв'язання поставленої задачі при задовільних часових характеристиках процедур навчання забезпечують ансамблеві методи за різними схемами побудови.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Lee, J. K. Artificial Intelligence Applications in Electronic Commerce. *PRICAI 2000 Topics in Artificial Intelligence*. Springer, Berlin, Heidelberg, 2000. Pp. 4–42.
2. Cheng D. The Research of Personalization E-Commerce Model Based on Data Mining. *International Conference on Management and Service Science.* 2011. Pp. 1–43.
3. Vysotska V., Chyrun L. Conceptual model of electronic content commerce systems. *Radio Electronics, Computer Science, Control*. 2014. № 1. Pp. 46–54.
4. Raghavan N-R. S. Data mining in e-commerce: A survey. *Sadhana*. 2005. Vol. 30. Is. 2–3. Pp. 275–289.
5. Karat C-M, Blom J. O., Karat J. Designing Personalized User Experiences in eCommerce. *Springer: Human–Computer Interaction Series*. Springer, Netherlands, 2004. P. 348.
6. Smith M., Wenerstrom B., Giraud-Carrier C., Lawyer S., Liu W. Personalizing E-Commerce with Data Mining. In: Lu J., Zhang G., Ruan D. (eds) *E-Service Intelligence. Studies in Computational Intelligence.* Springer, Heidelberg, 2007. Vol 37. Pp. 273–286.
7. Tkachenko R., Duriagina Z., Lemishka I. et al. Development of machine learning method of titanium alloy properties identification in additive technologies. *Eastern-European Journal of Enterprise Technologies*. 2018. № 3. Pp. 23–31.
8. Joshi R., Gupte R., Saravanan P. A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers. *Theoretical Economics Letters*. 2018. № 08. P. 448.
9. Wu X., Meng S. E-commerce customer churn prediction based on improved SMOTE and AdaBoost. *13th International Conference on Service Systems and Service Management (ICSSSM)*. Kunming, 2016. Pp 1–5.
10. Cao Y., Miao Q-C., Liu J-C., Gao L. Advance and Prospects of AdaBoost Algorithm. *Acta Automatica Sinica*. 2013. Vol. 39. Is. 6. Pp. 745–758.
11. Alomair O. A., Garrouch A. A. A general regression neural network model offers reliable prediction of CO2 minimum miscibility pressure. *Journal Petrol Explor Prod Technol.* 2016. № 6. Pp. 351–365.
12. Tkachenko R., Izonin I. Model and Principles for the Implementation of Neural-Like Structures Based on Geometric Data Transformations. In: Hu Z, Petoukhov S, Dychka I, He

M (eds) *Advances in Computer Science for Engineering and Education*. Springer International Publishing, Cham, 2019. Pp. 578–587.
13. Izonin I., Trostianchyn A. et al. The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production. *International Journal of Intelligent Systems and Applications*. 2018. 10:40–47.
14. Tepla T. L., Izonin I. V., Duriagina Z. A. et al. Alloys selection based on the supervised learning technique for design of biocompatible medical materials. *Archives of Materials Science and Engineering*. 2018. № 1. Pp. 32–40.
15. Black Friday Bonanza. 2019. URL: https://kaggle.com/mytymohan/black-friday-bonanza. (дата звернення 30.05.2019).
16. EDA + REGRESSION + CLASSIFICATION FROM SCRATCH. 2019. URL: https://kaggle.-com/rahu7292/eda-regression-classification-from-scratch. (дата звернення 30.05.2019).

REFERENCES

1. Lee, J. K. (2000). Artificial Intelligence Applications in Electronic Commerce. *PRICAI 2000 Topics in Artificial Intelligence*. Springer, Berlin, Heidelberg, 4–42 (in English).
2. Cheng, D. (2011). The Research of Personalization E-Commerce Model Based on Data Mining. *International Conference on Management and Service Science,* 1–43 (in English).
3. Vysotska, V., & Chyrun, L. (2014). Conceptual model of electronic content commerce systems. *Radio Electronics, Computer Science, Control*, 1, 46–54 (in English).
4. Raghavan, N-R. S. (2005). Data mining in e-commerce: A survey. *Sadhana*, 30, 2–3, 275–289 (in English).
5. Karat, C-M, Blom, J. O., & Karat, J. (2004). Designing Personalized User Experiences in eCommerce. *Springer: Human–Computer Interaction Series*. Springer, Netherlands, 348 (in English).
6. Smith, M., Wenerstrom, B., Giraud-Carrier, C., Lawyer, S., & Liu, W. (2007). Personalizing E-Commerce with Data Mining. In: Lu J., Zhang G., Ruan D. (eds) *E-Service Intelligence. Studies in Computational Intelligence.* Springer, Heidelberg, 37, 273–286 (in English).
7. Tkachenko, R., Duriagina, Z., & Lemishka, I. et al. (2018). Development of machine learning method of titanium alloy properties identification in additive technologies. *Eastern-European Journal of Enterprise Technologies*, 3, 23–31 (in English).
8. Joshi, R., Gupte, R., & Saravanan, P. (2018). A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers. *Theoretical Economics Letters*, 08, 448 (in English).
9. Wu, X., & Meng, S. (2016). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. *13th International Conference on Service Systems and Service Management (ICSSSM)*. Kunming, 1–5 (in English).
10. Cao, Y., Miao, Q-C., Liu, J-C., & Gao, L. (2013). Advance and Prospects of AdaBoost Algorithm. *Acta Automatica Sinica*, 39, 6, 745–758 (in English).
11. Alomair, O. A., & Garrouch, A. A. (2016). A general regression neural network model offers reliable prediction of CO2 minimum miscibility pressure. *Journal Petrol Explor Prod Technol,* 6, 351–365 (in English).
12. Tkachenko, R., & Izonin, I. (2019). Model and Principles for the Implementation of Neural-Like Structures Based on Geometric Data Transformations. In: Hu Z, Petoukhov S, Dychka I, He

M (eds) *Advances in Computer Science for Engineering and Education*. Springer International Publishing, Cham, 578–587 (in English).
13. Izonin, I., & Trostianchyn, A. et al. (2018). The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production. *International Journal of Intelligent Systems and Applications*, 10:40–47 (in English).
14. Tepla, T. L., Izonin, I. V., & Duriagina, Z. A. et al. (2018). Alloys selection based on the supervised learning technique for design of biocompatible medical materials. *Archives of Materials Science and Engineering*, 1, 32–40 (in English).
15. Black Friday Bonanza. (2019). Retrieved from https://kaggle.com/mytymohan/black-friday-bonanza. (30.05.2019) (in English).
16. EDA + REGRESSION + CLASSIFICATION FROM SCRATCH. (2019). Retrieved from https://kaggle.com/rahu7292/eda-regression-classification-from-scratch (30.05.2019) (in English).

doi: 10.32403/1998-6912-2019-1-58-62-70

THE RESEARCH AND EXPERIMENTAL ANALYSIS OF MACHINE LEARNING METHODS IN E-COMMERCE TASKS

P. B. Vitynskiy[1], R. O. Tkachenko[1], B. M. Havrysh[2]

[1]Lviv Polytechnic national University,
12, S. Bandera St., Lviv, 79013, Ukraine

[2]Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine
pavlo.vitynsky@gmail.com,
roman.tkachenko@gmail.com

The development of commercial activity by means of information systems provides many advantages for consumers of services and for companies that make such proposals. Exploring new markets for goods and services by internet companies provides great opportunities to consumers for analyzing and buying new products. The rapid development of computer technologies and information systems allows customers to buy the product even on the phone. However, the part of retail sales using such systems is still very small. A traditional e-commerce model based on finding a product or service on the Internet is becoming a bottleneck, impeding further company development. Instead, it can be replaced by personalized systems that give many benefits.

The problem of big data analytics in personalized systems of electronic commerce has been considered. An overview of existing methods of machine learning (Random Forest, AdaBoost, Multilayer Perceptron, SGTM neural-like structure, Linear Regression using Gradient Descent, General Regression Neural Network, Regressor Based on the Support Vectors Machine) for solving regression problems in such systems has been made. The advantages and disadvantages of the considered methods and algorithms

have been presented. The methods of machine learning have been used for forecasting the number of consumer expenses of the retail store. The accuracy of the forecasting and the time of training procedures of all considered methods has been experimentally determined. Despite the high temporal characteristics of the work, methods based on the Support Vectors Machine and the Stochastic Gradient Descent do not provide sufficient precision. The efficiency of neural network methods is also not satisfactory in terms of accuracy and training time. It has been established that the ensemble methods provide the highest accuracy with satisfactory time characteristics of the training procedures for a given problem.

Keywords: *ensemble, e-commerce, machine learning, forecasting, random forest, regression task, artificial intelligence.*

Стаття надійшла до редакції 03.04.2019
Received 03.04.2019