УДК 004.67

ОГЛЯД МЕТОДІВ ТА ЗАСОБІВ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ БАНКРУТСТВА КОМПАНІЙ

Федоренко Ілля, Луньова Марія

Науковий керівник: доктор філософії з Прикладної математики, старший викладач кафедри математики та цифрових технологій Луньова М.В.

Центральноукраїнський державний університет імені Володимира Винниченка, м. Кропивницький, Україна

Наукова стаття присвячена важливості дослідження та використання великих даних (Big Data) у сучасних інформаційних технологіях. Вона охоплює основні характеристики великих даних, такі як обсяг, швидкість, різноманітність, достовірність, мінливість, візуалізація та цінність, та висвітлює труднощі зберігання і обробки цих даних. Стаття акцентує увагу на машинному навчанні, висвітлюючи різні методи машинного навчання, такі як лінійна регресія, логістична регресія, дерева рішень, випадковий ліс та k-найближчих сусідів, описуючи їхні особливості та застосування у класифікації та регресії. Наприкінці статті надається короткий огляд популярних аналітичних систем, таких як RapidMiner, KNIME, Weka та ін., які використовуються для обробки та аналізу великих наборів даних.

Ключові слова: *великі дані, машинне навчання, контрольоване навчання.*

Review of methods and tools of machine learning for the analysis and forecasting of company bankruptcy

I. Fedorenko, M. Lunova

Academic supervisor: Doctor of Philosophy in Applied Mathematics, senior lecturer of the Department of Mathematics and Digital Technologies Lunova M.V.

Volodymyr Vynnychenko Central Ukrainian State University, Kropywnytsky, Ukraine

The scientific article is devoted to the importance of research and use of Big Data in modern information technologies. It covers the main characteristics of Big Data, such as volume, velocity, variety, reliability, variability, visualization and value, and highlights the challenges of storing and processing this data. The article focuses on machine learning, highlighting various machine learning techniques such as linear regression, logistic regression, decision trees, random forest, and k-nearest neighbors, describing their features and applications in classification and regression. At the end of the article, a brief overview of popular analytical systems such as RapidMiner, KNIME, Weka, etc., which are used to process and analyze Big Data sets, is provided.

Keywords: *big data, machine learning, supervised learning.*

Постановка проблеми. У сучасному світі об'єми доступної інформації про економічну діяльність компаній надзвичайно великі, а їх аналіз стає важливою складовою прийняття управлінських рішень та визначення стратегій розвитку. Засоби машинного навчання надають можливість ефективно моделювати та прогнозувати великі обсяги даних, що робить їх потужним інструментом у сфері фінансового аналізу.

Особливо актуальною стає задача прогнозування банкрутства компаній, оскільки це дозволяє уникнути фінансових криз та вчасно реагувати на ризики. Здійснення прогнозів на основі великих наборів економічних показників вимагає використання новітніх методів аналізу та інтеграції засобів машинного навчання.

Дана стаття спрямована на дослідження можливостей та ефективності використання методів машинного навчання для створення моделей, що передбачають фінансові труднощі та банкрутство компаній на основі комплексного аналізу їхньої економічної діяльності. Побудова таких моделей може служити основою для розробки ефективних стратегій фінансового управління та підвищення стійкості бізнесу в умовах зростаючої економічної невизначеності. Такий підхід до прогнозування фінансових труднощів є важливим етапом для підтримки сталого розвитку підприємств та їхньої взаємодії з фінансовими інституціями.

Аналіз досліджень і публікацій. Стрімкий розвиток технологій та інтернет-зв’язку призвів до вибуху кількості даних, що генеруються в усьому світі. За оцінками, станом на 2019 рік людство продукувало 2,5 квінтильйона байт даних на день [1]. Цей величезний обсяг даних, відомий як великі дані, трансформував індустрії та відкрив нові сфери можливостей. Однак використання великих даних у значущий спосіб залишається постійним викликом.

Актуальність великих даних зумовлена їхнім потенціалом розкривати натуру явищ та сприяти прийняттю більш ефективних рішень у різних сферах. Застосовуючи аналітику і машинне навчання до величезних, різнорідних наборів

даних, організації можуть виявити раніше приховані закономірності і взаємозв'язки. Це може призвести до конкурентних переваг, інновацій та оптимізації процесів [2]. Наприклад, великі дані уможливлюють персоналізовану охорону здоров'я, поєднуючи клінічні дані з генетичними даними пацієнта та даними про спосіб життя [3]. Ритейлери оптимізують ціноутворення, запаси та рекомендації, аналізуючи дані про точки продажу, ланцюги поставок та клієнтів [4].

Мета статті: огляд методів та засобів машинного навчання для аналізу та моделювання великих наборів даних діяльності компаній.

Виклад основного матеріалу (результатів) дослідження. Великі дані (Big Data) в інформаційних технологіях – це структуровані та/або неструктуровані набори інформації настільки великих розмірів, що традиційні засоби та підходи не можуть бути застосовані до них [5]. Також, що в залежності від контексту, під цим поняттям можуть розуміти і деякий набір інструментів та методів (наприклад, засоби обробки та зберігання даних, зокрема системи NoSQL, алгоритми MapReduce, чи програмний каркас проекту Hadoop).

Однак реалізація потенціалу великих даних не позбавлена труднощів. Існує безліч характеристик для великих даних, розглянемо основні з них [5]:

Volume (обсяг). Обсяг даних, який невпинно зростає, (очікується, що до 2025 року обсяги даних можуть досягти 181 зетабайт), вкотре підкреслює необхідність спеціальних технологій для їхнього аналізу.

Velocity (швидкість). Виклик Big Data полягає в ефективному вирішенні завдань зі швидкістю оновлення даних, щоб аналіз залишався актуальним у режимі реального часу.

Variety (різноманітність). Розмаїття форматів даних відзначає необхідність індивідуального підходу до аналізу, охоплюючи структуровані та неструктуровані дані, текст, графіку та інші джерела.

Veracity (достовірність). Важливість достовірності даних для уникнення неправильних управлінських рішень та невдач у реалізації стратегій.

Variability (мінливість). Потреба постійного відстеження змін в контексті даних, щоб адаптувати стратегії до різноманітних обставин.

Visualization (візуалізація). Використання якісної візуалізації даних, що включає зрозумілі діаграми та інтерактивні елементи, для полегшення сприйняття результатів аналізу.

Value (цінність). Важливо не лише аналізувати великі дані, але й мати можливість витягти максимум корисної інформації з отриманих результатів для ефективного управління та конкурентної переваги.

Управління та обробка таких величезних, швидкозмінних, неструктурованих даних вимагає передових аналітичних навичок і масштабованої інфраструктури зберігання та обчислень. Такі питання, як якість даних, конфіденційність, безпека та етичне використання алгоритмів також вступають у гру. Інтеграція великих даних у бізнес-процеси та перетворення аналізу на практичні дії залишаються постійними викликами [6].

Машинне навчання зробило революцію в багатьох галузях, дозволивши комп'ютерам навчатися на основі даних без явного програмування. Ключовою метою є розробка алгоритмів, які можуть отримувати вхідні дані та використовувати статистичний аналіз для виявлення закономірностей, які потім можуть бути використані для прогнозування або прийняття рішень у майбутньому [7]. Існує кілька основних підходів до машинного навчання (рис.1): контрольоване (навчання з учителем), неконтрольоване (навчання без учителя) та навчання з підкріпленням.

Рис. 1. Методи машинного навчання

Алгоритми *контрольованого навчання* будуються за допомогою маркованих наборів даних, які включають як вхідні дані, так і бажані результати [8]. Алгоритм аналізує ці пари вхідних-вихідних даних, щоб навчити модель, яка може робити прогнози для нових немаркованих даних.

Контрольоване навчання зазвичай використовується для таких задач:

- задача класифікації, яка полягає в тому, щоб допомогти розділити дані на різні категоріальні класи;

- задача регресії, яка полягає в тому, щоб показати або передбачити взаємозв'язок між процесом та тим, що цей процес може спровокувати.

Тобто, алгоритми контрольованого навчання застосовуються там, де доступні марковані навчальні дані, такі як виявлення спаму, розпізнавання зображень і прогнозування банкрутства компаній.

Неконтрольоване навчання аналізує немарковані набори даних для пошуку прихованих шаблонів, угруповань і взаємозв'язків у даних [9]. Методи, такі як кластеризація, зменшення розмірності, вивчення асоціативних правил та виявлення аномалій, надають можливість виявити приховані зв'язки і структури в наборах даних. Ці методи відкривають нові можливості для отримання цінної

інформації та використання її у практичних застосуваннях з невеликою кількістю попередньо встановлених категорій.

Навчання з підкріпленням є окремим випадком контрольованого навчання, що передбачає розробку агентів, які можуть навчитися оптимальній поведінці через взаємодію з навколишнім середовищем методом спроб і помилок [10]. Агент отримує винагороду або покарання за дії, які впливають на його навчання. Фактично агент і середовище утворюють систему зі зворотним зв'язком. Даний підхід у навчанні використовується, наприклад, у системах навігації для роботів, в логістиці, при складанні графіків і плануванні завдань.

Для аналізу та прогнозування банкрутства компаній, на нашу думку, найкраще підходить контрольоване навчання. Це пов’язано з тим, що для цих задач доступні дані з відповідними маркерами, зокрема дані про фінансові показники компаній. Контрольоване навчання також є хорошим вибором, оскільки воно дозволяє створювати моделі з достатньо високим показником точності та надійності. Це важливо для задач прогнозування банкрутства, оскільки помилки в прогнозах можуть призвести до значних фінансових втрат.

Однак, якщо немає доступу до даних з відповідними маркерами, то можна використовувати навчання без учителя або навчання з підкріпленням. Наприклад, можливо використовувати навчання без учителя, щоб навчити модель класифікувати компанії на такі, що збанкрутували і такі, що – ні.

До основних методів контрольованого машинного навчання можна віднести: метод k найближчих сусідів, регресійні моделі, байєсівські класифікатори, дерева рішень, випадковий ліс, нейронні мережі (глибоке навчання).

Лінійна Регресія. Даний метод полягає у пошуку лінії найкращого підходу, яка може допомогти у прогнозуванні результату для безперервних залежних змінних. В основі методу лежить лінійна функція обчислення зв’язку між однією незалежною та однією залежною змінною. У множинній лінійній регресії є більше двох незалежних змінних.

Логістична Регресія. Даний метод полягає в тому, щоб класифікувати результати, які можуть бути лише в межах від 0 до 1. В основі методу лежить сигмоподібна функція, яка перетворює зважену суму вхідних даних на значення від 0 до 1. Логістична регресія використовується для прогнозування категоріальної залежної змінної з використанням інформації від незалежних факторів.

Дерево Рішень. Даний метод полягає у розділенні даних на окремі частини та використовує візуальні уявлення, щоб показати очікувані результати дій, витрати та наслідки. Основна мета методу — побудувати модель, яку можна використовувати для прогнозування класу цільової змінної.

Випадковий Ліс. Даний метод відноситься до ансамблевих методів машинного навчання і полягає у тому, що для обробки завдань класифікації даний метод використовує категоріальні змінні, а для вирішення завдання регресії — використовує набори даних, які містять безперервні змінні.

K-найближчі сусіди. Даний метод полягає у визначенні відстані між точками даних, щоб позначити невидимі дані на основі найближчих позначених спостережуваних точок. Для визначення такої відстані можна використовувати такі вимірювання, як евклідова відстань, відстань Хеммінга, відстань Манхеттена та відстань Мінковського. Застосовується метод K-найближчих сусідів до задач класифікації та регресії.

Загалом, вибір правильного підходу до машинного навчання вимагає аналізу проблеми, характеристик даних і цілей продуктивності. Завдяки різноманітному набору алгоритмів і правильному налаштуванню машинне навчання може забезпечити трансформаційний вплив у різних галузях.

Сьогодні розроблено різноманітні інструменти та платформи, які дозволяють організаціям обробляти, аналізувати та вилучати цінність з великих гетерогенних наборів даних. Наведемо короткий огляд деяких широко розповсюджених аналітичних систем.

RapidMiner - це платформа для аналізу даних, що забезпечує інтегроване середовище для машинного навчання, предиктивної аналітики та бізнес-

аналітики [11]. Вона використовує графічний інтерфейс користувача з компонентами перетягування, що дозволяє розробляти аналітичні робочі процеси без програмування. Користувачі можуть отримати доступ до різних алгоритмів, джерел даних і перетворень даних у RapidMiner і комбінувати їх між собою. Розширюваність забезпечується завдяки інтеграції з мовами програмування, такими як Python та R.

Orange - це інструмент візуалізації та аналізу даних з відкритим вихідним кодом з візуальною побудовою робочого процесу програмування [12]. Графічний інтерфейс дозволяє створювати робочі процеси з інтелектуального аналізу даних та машинного навчання, збираючи попередньо упаковані віджети, які не потребують кодування. Основні можливості включають попередню обробку даних, розробку функцій, оцінку моделей та інтерактивну візуалізацію даних. Orange інтегрує бібліотеки Python, такі як NumPy, SciPy, Scikit-Learn, та підтримує додаткові бібліотеки.

KNIME (Konstanz Information Miner) - це модульний аналітичний фреймворк з відкритим вихідним кодом, що забезпечує візуальну збірку конвеєрів даних [13]. Інтуїтивно зрозумілий інтерфейс drag-and-drop дозволяє комбінувати кроки обробки даних та виконувати окремі етапи аналізу для повторного використання. Код на Python, R та Java може бути інтегрований у робочі процеси KNIME. Він широко застосовується в таких галузях, як бізнес-аналітика, фармацевтичні дослідження та аналіз споживачів.

Weka - це ветеранський програмний інструментарій машинного навчання з відкритим вихідним кодом, розроблений в Університеті Вайкато, Нова Зеландія [14]. Він включає в себе інструменти для візуалізації даних, попередньої обробки, класифікації, кластеризації, регресії та видобування правил асоціацій. Weka дозволяє розробляти нові алгоритми машинного навчання на Java і підключається до зовнішніх бібліотек Java.

Alteryx Analytics - це аналітична платформа з низьким рівнем/без коду для змішування, підготовки та аналізу даних [15]. Вона дозволяє підключатися до понад 140 джерел даних, включаючи файли, бази даних та хмарні додатки.

Alteryx спрощує створення та спільне використання повторюваних робочих процесів. Він надає можливості предиктивного, просторового та текстового аналізу. Хмарна версія уможливлює співпрацю між аналітиками даних та науковцями.

Таким чином, ці платформи та інші інструменти, такі як SPSS, MATLAB, SAS тощо, надають потужні можливості для отримання ефективної інформації та цінності з великих даних у різних сферах використання.

Висновки та перспективи подальших пошуків у напрямі дослідження. Швидкість зростання обсягів інформації у різних галузях відкрила нові можливості для отримання цінних результатів через передову аналітику. Проте, використання великих даних вимагає вирішення проблем, таких як їх обсяг, різноманітність, швидкість, а також питання конфіденційності, безпеки та етики.

Машинне навчання стає ключовим інструментом для виявлення закономірностей у великих та різноманітних наборах даних. Відповідні методи контрольованого, неконтрольованого навчання, а також навчання з підкріпленням дозволяють вирішувати різноманітні проблеми в залежності від наявних даних і поставлених цілей. Вибір відповідних алгоритмів та моделей може значно підвищити продуктивність.

Для ефективного використання великих даних з'явилися різноманітні інструменти та платформи, такі як RapidMiner, KNIME, Weka, Alteryx, а також система Python з відкритим вихідним кодом. Проте, для досягнення користі від великих даних необхідне створення міждисциплінарних команд, що поєднують знання предметної області, аналітичні навички та бізнес-орієнтацію.

Щодо подальших досліджень у цьому напрямку, важливим є розгляд можливостей застосування алгоритмів машинного навчання для аналізу діяльності компаній та побудови моделей прогнозування їхнього банкрутства.

Список використаної літератури

1. Desjardins, J. (2019). How Much Data is Generated Each Day?. Visual Capitalist. https://www.visualcapitalist.com/how-much-data-is-generated-each-day/

2. McAfee, A. & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review. https://hbr.org/2012/10/big-data-the-management-revolution

3. Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 1-10. https://doi.org/10.1186/2047-2501-2-3

4. Erevelles, S., Fukawa, N. & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897-904. https://doi.org/10.1016/j.jbusres.2015.07.001

5. Що таке Big Data? Джерело: https://hub.kyivstar.ua/articles/shho-take-big-data (01 листопада 2023)

6. Kwon, O., Lee, N. & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International Journal of Information Management, 34(3), 387-394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

7. Murphy, K.P. (2022). Probabilistic Machine Learning: An introduction. MIT Press.

8. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning. MIT Press

9. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer.

10. Sutton, R.S., & Barto, A.G. (2018). Reinforcement Learning: An Introduction. 2nd ed. MIT Press.

11. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 935-940).

12. Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L.,

Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in python. Journal of Machine Learning Research, 14, 2349-2353.

13. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... & Wiswedel, B. (2008). KNIME: The Konstanz information miner. In Data analysis, machine learning and applications (pp. 319-326). Springer, Berlin, Heidelberg.

14. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

15. Alteryx (2021). The Alteryx Analytic Process: Data science meets business intelligence. https://www.alteryx.com/resources/research-reports/analytics-revolution-research-summary