2. Жаргалсайхан Н. Особенности цифровой трансформации зарубежных компаний: анализ опыта компании General Electric. *Стратегии бизнеса. Электронный научно-экономичный журнал*. Том 9. № 2 (2021). С. 42-48.

3. Китова О.В., Брускин С.Н. Цифровая трансформация бизнеса. *Цифровая экономика*. 1(1). ‘2018. С. 20-25.

4. Дергачова Г. М., Колешня Я. О. Цифрова трансформація Бізнесу: сутність, ознаки, вимоги та технології. *Економічний Вісник НТУУ «КПІ»*. 2020. С. 280-290.

5. Перспективные технологии. https://home.kpmg/ru/ru/home/services/consulting/technology-services/emerging-technologies-in-risk-consulting.html

6. Susan Moore. These data and analytics technology trends will have significant disruptive potential over the next three to five years. https://www.gartner.com/smarterwithgartner/gartner-top-10-data-analytics-trends/.

7. Kasey Panetta. The Gartner Hype Cycle for Emerging Technologies, 2020 highlights 30 technology profiles that will significantly change society and business over the next five to ten years. https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020.

Статтю подано до редакції 17.10.2021

УДК 519.868:339.92 DOI 10.33111/mise.101.14

Юнькова О.О., к. ф.-м. н., доцент,
професор кафедри математичного моделювання та статистики,
ДВНЗ «КНЕУ імені Вадима Гетьмана»
Володько Т.О., магістрант 1 курсу
за освітньо-професійною програмою
«Економічна кібернетика і Дата Сайнс», ІІТЕ,
ДВНЗ «КНЕУ імені Вадима Гетьмана»

Yunkova O.O., PhD in Mathematics,
Professor of the Department of Mathematical Modeling and Statistics,
SHEI KNEU named after V. Hetman
Volodko T.O.,
1th grade Master’s student for the educational
and professional program “Economic Cybernetics and Data Science”
of the Institute of Information Technology in Economics,
SHEI KNEU named after V. Hetman

ПРОГНОЗУВАННЯ ПОПУЛЯРНОСТІ ІНТЕРНЕТ-КУРСІВ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

PREDICTING THE POPULARITY OF INTERNET COURSES BY MACHINE LEARNING METHODS

Аннотація. *Масові відкриті онлайн-курси (МВОК) — приклад розвитку руху відкритого навчання, яке привернуло велику увагу як академічної, так і громадської сфери. МВОК не є самостійним явищем, ізольованим від інших*

розробок в області відкритого і дистанційного навчання або освітніх технологій. Навпаки, МВОК тісно пов'язані з іншими розробками в цій галузі, мають потенціал для підтримки навчання протягом усього життя, усунення перешкод у процесі навчання, забезпечення рівності можливостей в освіті і, що найголовніше, забезпечення лібералізації знань. У роботі визначено теоретичні засади формування ринку інтернет-курсів; проаналізовано сучасний стан і тенденції ринку інтернет-курсів України та світу; проведена класифікація інтернет-курсів залежно від їх рейтингової оцінки; прогнозується значення рейтингової оцінки для визначення популярності онлайн-курсів. Вирішення проблеми прогнозування популярності курсів у даному дослідженні досягається завдяки методам машинного навчання, які класифікують онлайн курси на основі параметру рейтингової оцінки. А саме, курси, які отримують максимальний рейтинговий бал, вважаються популярними.

Розв'язування задач класифікації чи регресії засобами машинного навчання найчастіше досягається шляхом побудови ансамблевих моделей. В основу такого підходу покладено гіпотезу про об'єднання кількох моделей, яке може призвести до утворення потужнішої моделі. Спосіб об'єднання моделей має бути адаптованим до їхніх типів. Наразі існує кілька мета-алгоритмів, що застосовують для утворення об'єднаних моделей. В одному з них (метод беггінгу) однорідні початкові моделі навчаються паралельно та незалежно одна від одної, а потім об'єднуються згідно певного детермінованого правила усереднення. Одним з варіантів цього алгоритму є метод випадкового лісу. В іншому алгоритмі моделі навчаються послідовно в адаптивний спосіб. Найпопулярні з них — адаптивний і градієнтний бустинг. Перший оновлює вагу кожного з об'єктів навчального датасета, а другий — оновлює значення цих об'єктів. При цьому обидва методи намагаються розв'язати задачу оптимізації для пошуку найкращої моделі, представленої зваженою сумою початкових слабших моделей. У даній роботі для прогнозування популярності інтернет-курсів застосовано алгоритми градієнтного бустингу та випадкового лісу. Запропоновані моделі гарантують 65-ти відсоткову точність прогнозів. Серед факторів, що знижують точність прогнозу-вання, можна назвати атрибути, які не дуже корелюють із прогнозним значенням, а також диспропорція та значні викиди, які спостерігаються у даних. Розглянуті методи машинного навчання піддаються модифікаціям та тюнінгу, що дає можливість покращити моделювання класифікатора.

Ключові слова: *інформаційні технології, інтернет-навчання, алгоритм градієнтного бустингу, алгоритм випадкового лісу*

Abstract. *Mass open online courses (MOOC) are an example of the development of the open learning movement, which has attracted a lot of attention from both the academic and public spheres. IOC is not an independent phenomenon isolated from other developments in the field of open and distance learning or educational technologies. On the contrary, IOCs are closely linked to other developments in this field, have the potential to support lifelong learning, remove barriers to learning, ensure equal opportunities in education and, most importantly, liberalize knowledge.*

The theoretical bases of formation of the market of Internet courses are defined in the work; the current state and trends of the Internet courses market in Ukraine and the world are analyzed; the classification of Internet courses depending on their rating assessment is carried out; the rating value is predicted to determine the popularity of online courses. The solution to the problem of predicting the popularity of courses in this study is achieved through machine learning methods that classify online courses based on the rating parameter. Namely, the courses that receive the maximum rating score are considered popular. Solving problems of classification or regression by machine learning is most often achieved by building

ensemble models. This approach is based on the hypothesis of combining several models, which could lead to the formation of a more powerful model. The method of combining models should be adapted to their types. Currently, there are several meta-algorithms used to form integrated models. In one of them (the method of bugging) homogeneous initial models are studied in parallel and independently of each other, and then combined according to a certain deterministic averaging rule. Currently, there are several meta-algorithms used to form integrated models. In one of them (the method of bugging) homogeneous initial models are studied in parallel and independently of each other, and then combined according to a certain determined averaging rule. One variant of this algorithm is the random forest method. In another algorithm, models are trained sequentially in an adaptive manner. The most popular of these are adaptive and gradient boosting. The first updates the weight of each of the training dataset objects, and the second updates the values of these objects. In doing so, both methods attempt to solve the optimization problem to find the best model represented by the weighted sum of the initial weaker models. In this paper, gradient boosting and random forest algorithms are used to predict the popularity of online courses. The proposed models guarantee 65 percent accuracy of forecasts. Factors that reduce the accuracy of the forecast include attributes that do not correlate much with the forecast value, as well as the disparity and significant emissions observed in the data. The considered methods of machine learning are subject to modifications and tuning, which makes it possible to improve the modeling of the classifier.
Keywords: *information technologies, e-leaming, gradient boosting algorithm, random forest algorithm*

Актуальність: У 21-му столітті відбулася зміна освітньої парадигми, пов'язана з широким використанням інформаційно-комунікаційних технологій (ІКТ). З поширенням ІКТ у мережі Інтернет відкрите та гнучке навчання перемістилося з периферії освітньої діяльності до основного напрямку освіти. ІКТ покращили якість і можливості онлайн-доставки освітнього контенту. Онлайн-мережі зараз широко використовуються як розподілені, гнучкі та доступні навчальні середовища і, що головне, потенційно відкриті.

Актуальність прогнозування популярності онлайн-курсів незаперечна, оскільки пандемія COVID-19 привернула багатьох людей до онлайн-освіти. Третина учнів, які коли-небудь реєструвались на платформі масових відкритих онлайн-курсів — МВОК, приєдналися до них саме у 2020 році. Масові відкриті онлайн-курси стали альтернативою освітньої платформи, яка дає змогу отримати доступ до такої ж якості навчання через Інтернет учням з віддалених географічних територій.

Аналіз останніх досліджень і публікацій. Розвиток машинного навчання характеризується появою складних і вдосконалених алгоритмів для розв'язування, здавалося, звичайних задач. Ці алгоритми пропонують нові підходи до роботи із даними, детально аналізують їх, підвищуючи тим самим статистичну значущість прогностичної моделі.

Прикладом таких підходів є алгоритми, побудовані на ідеї створення ансамблів моделей, відомі як алгоритми бустингу. Побудова ансамблів моделей базується на ітераційному перенавчанні, коли за допомогою простої моделі, яка сама по собі мала б нерелевантні прогнозні значення, будується потужніша за своїми характеристиками модель. Навчання моделі відбувається в такий спосіб, щоб кожна наступна модель підсилювала дію попередньої [1]. Найпопулярнішими серед подібних алгоритмів є алгоритм адаптивного бустингу та градієнтного бустингу [2].

Для покращення стабільності і точності алгоритмів машинного навчання застосовують випадковий ліс — модифікацію алгоритму дерева рішень, в якому замість побудови одного дерева будується кілька. Кожне з побудованих дерев виводить певний результат, а в якості остаточного результату обирається той, що зустрічається найчастіше [3].

Традиційні методи оцінки моделей, такі як коефіцієнт детермінації, скоригований коефіцієнт детермінації, коефіцієнт кореляції тощо, переважно не застосовують у задачах класифікації. Тут використовуються матриці помилок, показник accuracy, out-of-bag error і багато інших.

Інтерес до методів машинного навчання відображений у працях вітчизняних науковців Дербенцева В.Д, Жебки В. В., Бідюка П.І. Проблемам формування та розвитку ринку онлайн-освіти та інтернет-курсів в Україні присвячено низку наукових праць Н. В. Казаринової, Б. І. Шуневича, Г. Яценка, С. В. Степаненка, В. Ю. Стрельнікова та багатьох інших.

Невирішені проблеми. Не зважаючи на інтерес до даної теми, існує широке коло задач, які потребують вирішення сучасними методами машинного навчання. Однією з них є проблема визначення критеріїв, що впливають на популярність курсів серед студентів і напрямків, які є цікавими для сучасних учнів. У зв’язку зі стрімким розвитком онлайн-освіти кількість і різноманітність курсів, представлених на різних освітніх платформах дуже зросла: обрати курси за тематикою, змістом і викладачами стає дедалі тяжче. Тому було запропоновано оцінювати популярність онлайн-курсів на основі досліджуваних критеріїв.

Мета статті. Метою статті є аналіз ринку інтернет-курсів і дослідження факторів, що впливають на їхню популярність. Прогнозування популярності інтернет-курсів у таких умовах є важливим компонентом формування ринку електронної освіти та дає можливість організаторам інтернет-курсів керувати їхньою варті-

стю та своєю інвестиційною діяльністю, а студентам орієнтуватись на корисні та якісні знання.

Виклад основного матеріалу. Масові відкриті онлайн-курси (МВОК) — одна з найпомітніших тенденцій у вищій освіті останніх років. Термін «МВОК» транслює відкритий глобальний доступ до навчального контенту на основі відео, лекцій, завдань, тестів і форумів, які публікуються через онлайн-платформу для великої кількості учасників, які бажають пройти курс або здобути освіту [4, 5].

Кількість зареєстрованих користувачів МВОК постійно збільшується, так само збільшується і кількість курсів по всьому світу. Перший великий відкритий онлайн-курс, запроваджений Університетом Манітоби, припадає на 2008 рік. Однак широкий громадський інтерес до МВОК виявився після 2011 року, коли Стенфордський університет і Масачусетський технологічний інститут (MIT) поширили відкритий онлайн-курс зі штучного інтелекту. Цей курс охопив більше 160 000 студентів з більш ніж 190 країн [6].

Основними перевагами онлайн-курсів є: відкритість для всіх, незалежно від віку, місця проживання та соціального статусу; можливість вивчати різні дисципліни, не прив'язуючись до конкретної спеціалізації; безкоштовне використання; спілкування з однодумцями; швидкий пошук інформації; вся навчальна інформація зібрана в одному місці. Головною особливістю таких курсів можна вважати їхню інтерактивність, використання у будь-який зручний для користувача час [7]. Такий навчальний формат усуває фінансові обмеження, пов'язані з неможливістю дозволити собі отримувати освіту в університеті.

МВОК надає можливість студентам контролювати темп, в якому вони працюють. Кожен курс передбачає приблизний час, який потрібен на його проходження. І хоча самі курси обмежені у часі, немає жодних лімітів щодо перегляду навчального контенту протягом цього періоду. Перевагою використання курсів є те, що студенти з різних країн можуть взаємодіяти між собою, ділитись досвідом, разом провадити дослідження.

Складність контролю виконання завдань і ступеня залученості студентів є величезною проблемою, з якою стикаються як розробники курсів, так і інструктори. Одним із найсильніших аргументів проти МВОК є дуже низький рівень завершення курсів. Дослідження Кріса Парра та Кетті Джодан показали, що до завершення МВОК приходять лише 7 % слухачів, а частка їх починає знижуватися ще із першого тижня від запуску [8, 9]. Такі низькі

показники завершення курсів пов'язують з відсутністю персонального контакту та будь-якої форми взаємодії. Інша гіпотеза сформована навколо того, що власне акт завершення курсів не має значення, оскільки учні записуються туди, щоб знайти конкретну інформацію, а після її отримання зменшується необхідність витрачати свій час [10].

Спалах пандемії COVID-19 лише збільшив попит на МВОК по всьому світу. Ця тенденція, ймовірно, зберігатиметься і надалі, оскільки уряди багатьох країн виступають з ініціативами зробити освіту та навчання доступними для всіх. Очікується, що ринок онлайн-курсів отримає поштовх у найближчі роки [11]. При цьому одним із векторів розвитку буде створення та просування масових інформаційних платформ, і, зокрема, спеціальних платформ за напрямком навчання, що мають забезпечувати в одному місці доступ до величезного репозиторія даних із курсами [12].

Велика кількість інформації щодо студентів та їхньої поведінки, яку щоденно накопичують платформи відкритих масових онлайн курсів, дають змогу відслідковувати прогрес навчання та популярність курсів, алгоритми їх просування. Згенеровані та збережені дані з часом зростають експоненціально, але при цьому не приносять користі без використання спеціальних засобів аналізу, адже для ухвалення рішень необхідно класифікувати великі обсяги даних, відслідкувати залежності і робити прогнози на майбутні періоди.

При цьому дані можуть бути неповними або, навпаки, надмірними, зашумленими, не систематизованими, або систематизованими неправильно. Для усунення таких проблем використовують методи машинного навчання, які дають змогу варіювати параметри моделей в процесі навчання для досягнення найбільшої їх відповідності реальним даним. У зв'язку цим актуальним стає алгоритм адаптивного бустингу.

Алгоритм починає своє навчання з побудови дерева рішень, у якому на першій ітерації кожному значенню відповідає однакова вага. Після першого прогону алгоритму ті значення, які складно ідентифікувати та об'єднати у певний спосіб, отримують більше значення вагового коефіцієнту. На наступному етапі навчання алгоритму розпочинається зі значень, що мають найбільшу вагу, і у разі, якщо їх знову не вдалося класифікувати, то їхні коефіцієнти збільшуються. А для тих об'єктів, які класифікувати легше, значення коефіцієнтів зменшується.

Поряд із моделлю адаптивного бустингу розглядається модель градієнтного бустингу. В ній на першій ітерації для всього дата-

сету будується модель, на підставі якої виявляються помилки. На другій ітерації будується нова модель, що покликана їх мінімізувати з урахуванням цільової змінної. Прогнози моделей на наступних ітераціях базуються на знаннях, отриманих на попередніх кроках. Ітерації повторюються, допоки не буде досягнуто максимальної кількості класифікованих змінних або перестануть відбуватися зміни у значенні помилок.

Особливістю моделі XGBoost — різновиду моделі градієнтного бустингу — є наявна можливість тюнингу (коригування) параметрів поряд із дефолтними характеристиками. Процес тюнингу гіперпараметрів передбачає вибіркове присвоювання значень, або побудову сітки гіперпараметрів, серед яких алгоритм обирає їхню найкращу комбінацію.

Випадковий ліс — це одна з модифікацій алгоритму дерева рішень, що дає змогу суттєво підвищити точність класифікації. Метод базується на двох основних принципах: 1) у процесі тренування кожне дерево випадкового лісу вчиться на випадковому зразку з набору даних; 2) при поділі вузлів обираються випадкові набори параметрів, а для визначення найімовірнішого з них обирається той, що зустрічається найчастіше.

Цей алгоритм вважається одним з найуніверсальніших — його використовують у задачах регресії, кластеризації, селекції ознак, класифікації та пошуку аномалій.

Алгоритм дерева випадкового лісу має ряд переваг, які зумовлюють його широке використання, а саме:

1) випадковий ліс забезпечує значне підвищення точності за рахунок слабкої кореляції дерев в ансамблі, через подвійну ін'єкцією випадковості в індуктивний алгоритм — з аналізом помилок і використанням методу випадкових підпросторів при розщепленні кожної вершини;

2) усувається методологічно та алгоритмічно складна проблема скорочення повного дерева рішень, оскільки дерева випадкового лісу не скорочуються, що в свою чергу забезпечує високу обчислювальну ефективність;

3) усувається проблема перепідгонки навіть у разі більшої кількості ознак порівняно з кількістю спостережень навчальної вибірки і великої кількості дерев. Тим самим знімається складна проблема відбору ознак, що виникає в інших ансамблевих класифікаторів;

4. простота застосування: єдиними параметрами алгоритму є кількість дерев в ансамблі і кількість ознак, випадково відібраних для розщеплення у кожній вершині дерева;

5) ефективно обробляє нелінійні параметри: нелінійні параметри не впливають на продуктивність випадкового лісу, на відміну від інших алгоритмів;

6) простота організації паралельних обчислень [14].

Якість класифікації, що виконується алгоритмами машинного навчання, оцінюється на підставі матриці помилок, яка дає змогу виявляти статус класифікації значень. Модельні значення залежної змінної, отримані на навчальному наборі даних, порівнюють із реальними значеннями. Всім показникам призначається певний статус. При цьому для моделі бінарної класифікації уводиться статус «позитивний» для події зі значенням 1, а статус «негативний» для події зі значенням 0. Тоді після прогнозу для кожної із групи значень можна отримати два випадки: значення було класифіковано правильно і це ж значення було класифіковане неправильно, тобто потрапило не в ту групу. Залежно від комбінацій цих параметрів виокремлюють істинно-позитивні (*TP*) та хибно-позитивні значення (*FP*), істинно-негативні (*TN*) та хибно-негативні (*FN*) значення. Співвідношення таких значень зазвичай представлено у вигляді матриці (не обов'язково 2x2, це залежить від вхідних умов).

На основі значень матриці помилок визначається показник точності accuracy [15]:

$$accurancy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (1)$$

де *TP — кількість істинно-позитивних значень;*

TN — кількість істинно-негативних значень;

FP — кількість хибно-позитивних значень;

FN — кількість хибно-негативних значень.

Оцінити якість роботи моделі можна за допомогою показників точності (precision) та повноти (recall):

$$precision = \frac{TP}{TP+FP}, \quad (2)$$

$$recall = \frac{TP}{TP+FN}. \quad (3)$$

При цьому точність precision показує частку істино-позитивних значень, які були визначені як істино-позитивні, а повнота відображає, яку саме частину становлять істинно-позитивні об'єкти серед усіх позитивних об'єктів, які були класифіковані алгоритмом. За допомогою точності precision відбува-

ється певний розподіл об'єктів, оскільки якщо всі об'єкти віднести до одного класу, то отримаємо дуже високий рівень хибно-негативних значень.

Існує кілька підходів до об'єднання показників точності та повноти в один критерій. Зокрема, це може бути значення *F* — критерію як середньогармонічного між цими показниками, який бажано максимізувати [16]:

$$F_\beta = (1 + \beta^2) \frac{precision * recall}{(\beta^2 * precision) + recall}, \quad (4)$$

де β — *показник ваги точності*.

Іншим показником точності моделі є крос-ентопія (перехресна ентопія) або log loss — так звана функція витрат, яка використовується при розв'язанні задач логістичної регресії, а також при побудові моделей нейронних мереж.

Крос-ентопія є від'ємною логарифмічною імовірністю логістичної моделі. Для одного значення позитивно класифікованого об'єкта оцінка імовірності логарифмічних втрат буде такою:

$$L_{log}(y, p) = -(y(\log(p) + (1 - y)\log(1 - p)), \quad (5)$$

де p — *ймовірність приналежності до певного класу*;

y — *результативна змінна логістичної моделі.*

Слід зазначити, що в неперервному випадку цей показник має ту ж специфіку, що і поняття диференціальної ентропії, не зважаючи на формальну аналогію з дискретним випадком [17, 18].

У роботі розглядається набір даних «IT та програмне забезпечення», на платформі Udemy, що містить інформацію про близько 22 тисячі курсів, спрямованих на вивчення IT, програмного та апаратного забезпечення, мереж, безпеки даних, IT-сертифікації та інших напрямків [19]. Дослідження проведено в середовищі R.

Аналіз початкових даних показав, що кількість безкоштовних курсів щорічно зростала, а для категорії платних курсів у 2016 та 2020 роках відбулося незначне зниження в категорії «IT та програмне забезпечення». Це не означало зменшення популярності МВОК, а всього лише стало наслідком того, що курси публікувались на інших платформах або у інших категоріях. Крім того виявлено, що при постійному зростанні кількості онлайн-курсів та їхньої різноманітності, а також збільшенні кількості студентів, які долучаються до електронного навчання, середня кількість студентів на курс — спадала.

Детальніший аналіз даних стосувався атрибутів, які характеризують онлайн-курси. Всього їх налічується більше двадцяти.

Статистичну інформацію щодо змінних, а саме: мінімальні, максимальні та середні значення кожної числової змінної, їхні квантилі та моди, а також категоріальні та логічні змінні, отримано через функцію summary( ), а статистику щодо розподілу значень параметрів візуалізувано у вигляді «ящику з вусами» за допомогою функції boxplot( ).

Дані, які використовують для моделювання, мають відповідали вимогам алгоритмів. Насамперед їх потрібно очистити від зайвої інформації, перевірити, чи немає відсутніх значень, за необхідності закодувати певні атрибути.

З цією метою категоріальні змінні, які є унікальними для кожного запису, а також інші неінформативні дані, що не використовуються в моделюванні, або ж містять ідентичні дані, виключено із розгляду. Пропуски даних щодо окремих змінних заповнені за допомогою функції mutate( ). Використання функції dummyVars дало змогу закодувати логічні значення «True» в «1», а «False» у «0». Атрибути, що відповідають за дату створення і публікації курсів, замінено на два стовпці: «created_year» та «created_month» для «created» і «published_year» та «published_month» для «published_time».

У результаті отримано новий модифікований набір даних «udemy_tr», що складається з 22853 спостережень і має 12 числових атрибутів. Сила та напрям зв'язку між змінними виявляється за допомогою кореляційної матриці.

Популярність онлайн курсів великою мірою залежить від рейтингової оцінки курсу, пропонується класифікувати курси на 5 класів (табл. 1)

Для побудови моделі та її перевірки дані поділено на тестову («test») та навчальну («train») вибірки. При цьому до тренувального набору даних увійшло 15997 спостережень, тобто приблизно 70 % початкових даних, а до тестового 6856 — приблизно 30 % даних.

Таблиця 1

РЕЙТИНГОВА ОЦІНКА КУРСУ

Клас	Середній рейтинговий бал
перший	від 0 до 1
другий	від 1 до 2
третій	від 2 до 3
четвертий	від 3 до 4
п'ятий (найбільша популярність)	від 4 до 5, включно з 5

Для прогнозування популярність онлайн-курсів на платформі Udemy на основі відомих атрибутів використано алгоритм градієнтного бустингу.

При моделюванні показників із дефолтними параметрами алгоритм виконав 1000 прогонів і побудував стільки ж моделей. Вибір найкращої моделі здійснено за критерієм logloss — показника логарифмічної втрати. Мінімізація цього показника призводить до максимізації точності класифікатора.

Прогноз на основі алгоритму градієнтного бустинга з дефолтними параметрами показав результат точності 60,59 %.

Для підвищення точності прогнозування змінено дефолтні параметри: eta = 0.1, gamma = 1, max_depth = 5, min_child_weight = 3, subsample = 0.6, colsample_bytree = 0.8, а також максимальну глибину дерева, мінімальна суму ваги екземпляра, необхідну дочірнім атрибутам та інші характеристики. У результаті отримано модель, яка має точність класифікатора 63,46 %.

Для реалізації багатокласової класифікації застосовано алгоритм випадкового лісу (бібліотека library (randomForest)). Точність прогнозів за цим методом становить 64,82 %. З кращою точністю модель класифікувала курси, що належать до першого та п'ятого класів. Курси, з рейтинговою оцінкою від 1 до 2 та від 2 до 3 характеризуються низькою точністю прогнозів.

Для обох моделей вплив змінних є однаковим. А саме, на популярність онлайн-курс з категорії «IT та програмне забезпечення» на платформі Udemy, найбільший вплив має кількість відгуків і кількість учасників курсу. До того ж виявлена досить сильна кореляція між популярністю онлайн-курсів і кількістю опублікованих лекцій.

Найменш важливим показником для класифікації онлайн-курсів виявився параметр «is_paid_TRUE», що відповідає за те, чи є курс платним чи безкоштовним. Такий результат є несподіваним, оскільки безкоштовні курси мали б привертати увагу більшої кількості людей, однак з іншого боку, також можуть викликати недовіру до їхньої якості. Малий вплив цього параметр на прогнозні моделі, можна пояснити також «перекосом» даних, адже на 98 % оплачуваних курсів припадає лише 2 % безкоштовних.

Висновки. Аналіз масових відкритих онлайн курсів у категорії «IT та програмне забезпечення» на платформі Udemy показав, що ринок МВОК зростає з кожним роком як за кількістю створених курсів, так і за чисельністю нових студентів. На платформі значно переважають платні курси, хоча за останні 2 роки кількість опублікованих безкоштовних курсів також зросла. Пандемія

COVID-19 підсилила загальні тенденції, оскільки багато провайдерів і розробників курсів заохочують нових користувачів до таких курсів. З кожним роком конкуренція на ринку МВОК зростає: за останні 6 років кількість пропонованих курсів різними платформами зросла настільки, що середня кількість учасників курсів впала в кілька разів.

У результаті моделювання класифікаторів методами машинного навчання (градієнтним бустингом і випадковим лісом) отримано дві моделі, точність прогнозів яких становить 63,46 % і 65,5 % відповідно.

Відомо, що початкові дані визначають продуктивність системи машинного навчання. Якісний результат забезпечують якісні дані. Зокрема, перекоси даних, дублювання інформації, наявність неінформативних змінних можуть негативно впливати на результат моделювання.

У ході дослідження було з'ясовано, що на популярність онлайн-курсів насамперед впливають кількість відгуків, кількість лекцій у курсі та їхня ціна. Щодо напрямків курсів, то найпопулярнішими залишаються курси, які пов'язані із побудовою власного ІТ-бізнесу та машинним навчанням.

Бібліографічні посилання

1. Balabanov D. V., Kovtun A. V., Kravchenko Y. A. TWO-STAGE BOOSTING OF BINARY CLASSIFICATION BASED ON THE APPLICATION OF BIOINSPIRED ALGORITHMS. IZVESTIYA SFedU. *ENGINEERING SCIENCES*. 2020. No. 3. P. 133–146. URL: https://doi.org/10.18522/2311-3103-2020-3-133-146

2. Жебка В. В. Оптимізація роботи алгоритму градієнтного бустингу за допомогою перехресної перевірки [Електронний ресурс] / В. В. Жебка, В. І. Виноградов, А. П. Бондарчук, М. М. Степанов. *АКТУАЛЬНІ ПРОБЛЕМИ ЕКОНОМІКИ*. №12 (222). — 2019. — Режим доступу до ресурсу: https://ecoscience.net/archive/2019/APE-12-2019/12.19 topic_Zhebka %20VV, %20 Vynohradov %20VI, %20Bondarchuk %20A.P., %20Stepanov %20M.M..pdf.

3. Маслій Р.В. Застосування випадкових лісів для класифікації даних / Р.В. Маслій, О.Ю. Філіпчук. *Veda a technologie: krok do budoucnosti–2014*. 2014. Praha. Dfl. 30. — С. 24-27.

4. Massive open online courses (MOOCs) & Definitions — Educational Technology. Educational Technology. URL: https://educationaltechnology.net/massive-open-online-courses-moocs-definitions/

5. *eNUFTIR: Home*. URL: http://dspace.nuft.edu.ua/jspui/bitstream/123456789/19800/1/55.pdf.

6. Contributors to Wikimedia projects. Udemy — Wikipedia. Wikipedia, the free encyclopedia. URL: https://en.wikipedia.org/wiki/Udemy.

7. Шарова, Т. М. та Шаров, С. В. Масові відкриті онлайн курси як можливість підвищення конкурентоспроможності фахівця. *Молодий вчений*. 9.1 (61.1). 2018. С. 137-140. URL: http://eprints.mdpu.org.ua/id/eprint/2425/

8. Петренко С. В. Сутність та особливості українських платформ масових відкритих онлайн-курсів (МВОК). *Інноватика у вихованні*. 2020. Т. 2. № 11. С. 165–173. URL: https://doi.org/10.35619/iiu.v2i11.260

9. Parr Chris (2013), MOOC Completion Rates 'below 7 %': Open online courses' cohort much less massive at finish line, Retrieved from Times Higher Education on July 24th, 2015: https://www.timeshighereducation.co.uk/news/mooc-completion-rates-below-7/2003710.article

10. Jordan K. Massive Open Online Course Completion Rates Revisited: Assessment, Length and Attrition. *International Review of Research in Open and Distributed Learning.* 2015. Vol. 16, no. 3. P. 341–358. URL: https://files.eric.ed.gov/fulltext/EJ1067937.pdf

11. Glybovets M., Zhyrkova A. Using machine learning in sound classification tasks. NaUKMA Research Papers. *Computer Science*. 2019. Vol. 2. P. 22–31. URL: https://doi.org/10.18523/2617-3808.2019.2.22-31

12. 10 Predictions for the Online Course Industry in 2021. *Persuasion Nation: Passive Income Strategies for Busy People*. URL: https://www.persuasion-nation.com/blog/10-predictions-for-the-online-course-industry-in-2021

13. Massive Open Online Course Market Witnessing Impressive Growth Owing to the Surge in Demand amidst the Pandemic: FMI. URL: https://www.futuremarketinsights.com/press-release/massive-open-online-course-mooc-market

14. Чистяков С. П. Случайные леса: обзор. *Труды Карельского научного центра РАН*. 2013. № 1.С. 117–136. URL: http://resources.krc.karelia.ru/transactions/doc/trudy2013/trudy_2013_1_117-136.pdf.

15. Оценка классификатора (точность, полнота, F-мера). Суровая реальность. URL: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html.

16. Labintcev E. Метрики в задачах машинного обучения. Все публикации подряд / Хабр. URL: https://habr.com/ru/company/ods/blog/328372/.

17. sklearn.metrics.log_loss — scikit-learn 0.24.2 documentation. scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html.

18. Contributors to Wikimedia projects. Перекрёстная энтропия — Википедия. Википедия — свободная энциклопедия. URL: https://ru.wikipedia.org/wiki/Перекрёстная_энтропия.

19. IT & Software Courses Udemy — 22k+ courses. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/jilkothari/it-software-courses-udemy-22k-courses

20. By The Numbers: MOOCs in 2020 — Class Central. The Report by Class Central. URL: https://www.classcentral.com/report/mooc-stats-2020/.

Статтю подано до редакції 11.11.2021