УДК 004.94

Ю. Антюк, Т. Ліхоузова, Ю. Олійник

ПРОГНОЗУВАННЯ ПОПУЛЯРНОСТІ ОНЛАЙН-КУРСІВ НА ПЛАТФОРМІ COURSERA

Анотація: Робота присвячена побудові та дослідженню моделей прогнозування кількості студентів за основі рейтингу, кількості модулів, тривалості, рівня складності та типу розкладу курсів за допомогою лінійної та поліноміальної множинної регресії, Random forest та XGBoost та задачу класифікації курсів на популярні (більше 20000 студентів) та непопулярні (до 20000 студентів) методами логістичної регресії, Decision tree, Random forest та SVM. Результати роботи можуть бути корисними для людей, що шукають перевірені та популярні курси для навчання і для організацій, що є провайдерами курсів.

Ключові слова: інтелектуальний аналіз даних, модель прогнозування, модель класифікації.

Вступ

У сучасному світі, коли освіта стає все доступнішою, багато людей надають перевагу навчанню на онлайн-платформах. Онлайн-курси позиціонуються і сприймаються користувачами як якісний і зручний спосіб здобуття нових знань і навичок, який поступово стає все популярнішим. Такий формат освіти має низку переваг: можливість навчатися з будь-якої локації, гнучкий графік навчання, глобальний доступ до знань і багато іншого. Coursera [1] є однією з найпопулярніших платформ онлайн-курсів на сьогодні. Для аналітиків, викладачів та освітніх платформ важливим є прогноз популярності (кількості студентів) курсу. Це дозволяє покращувати навчальні програми та адаптувати контент до потреб слухачів.

Матеріали та методи

Для аналізу та моделювання було обрано 3 джерела відкритих даних на сайті Kaggle [2], що включають в себе:

− Найпопулярніші курси на платформі Coursera https://www.kaggle.com/datasets/elvinrustam/coursera-dataset

− Топ курсів в минулому (2024 році) https://www.kaggle.com/datasets/azraimohamad/coursera-courses-2024

− Датасет курсів з відгуками https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera?select=Coursera_reviews.csv

На основі детального опису та проведеного аналізу предметної області розроблено модель сховища даних курсів на платформі Coursera за типом „сніжинка” (рис.1). У моделі сховища спроектовано дві таблиці фактів, вісім таблиць вимірів та дві проміжні таблиці.

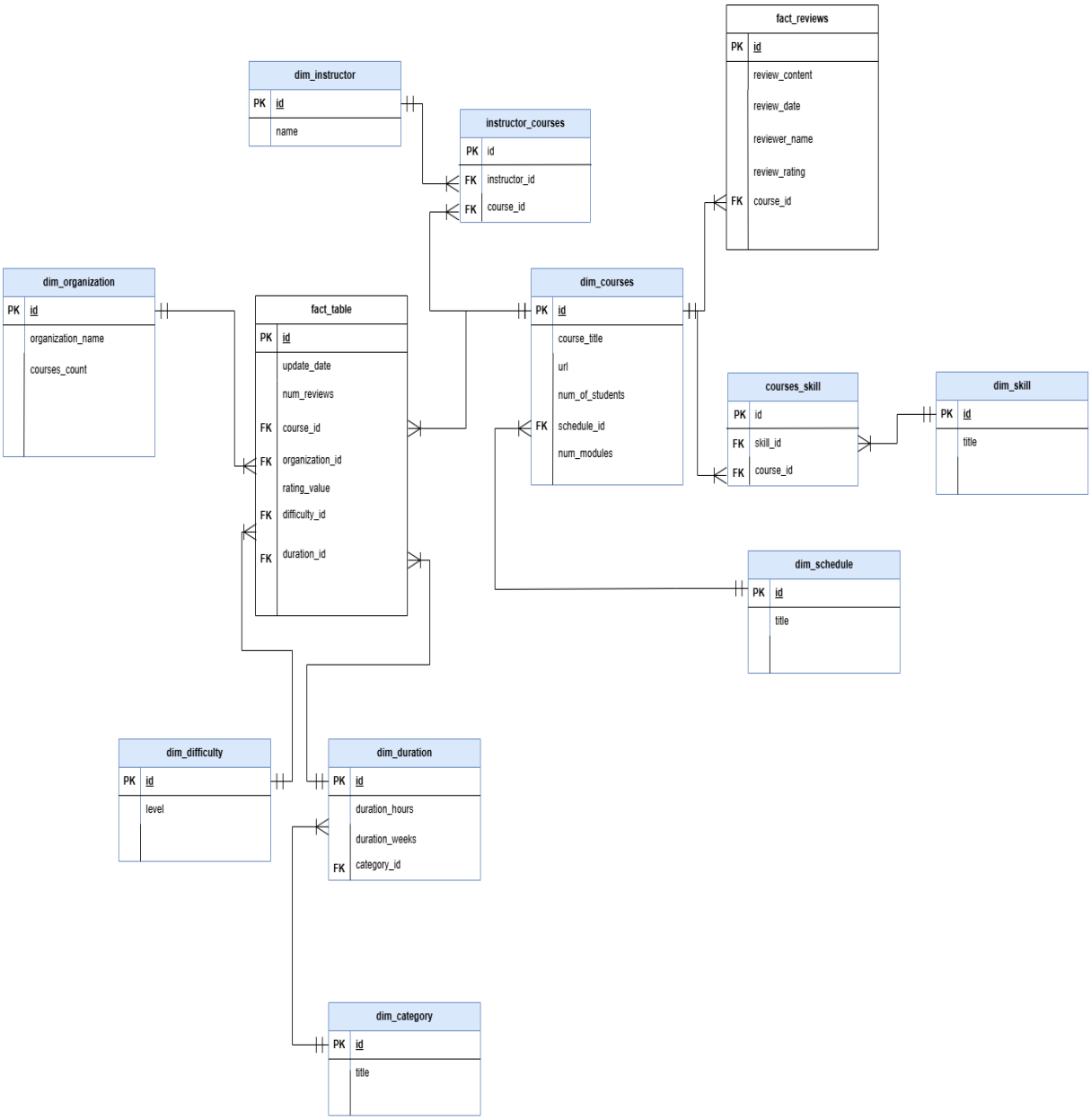

Рисунок 1. Модель сховища даних

Початкове дослідження даних

Лише чверть курсів має рейтинг менше 4.5. Найбільше курсів мають рейтинг 4.8 (22% всіх курсів). Найбільше курсів призначено для рівня складності beginner, а найменше – для advanced (рис.2). Це можна пояснити тим, що початківці частіше потребують додаткового навчання тоді, як для більш високого рівня знань може бути недостатньо експертів. Найдовша середня тривалість курсів для рівня advanced (близько 19 годин) тому, що часто такі курси покривають більш складні теми, на вивчення яких необхідно більше часу. Найбільше курсів навчають аналізу даних, програмуванню на Python та машинному навчанню.

 ISSN 1560-8956

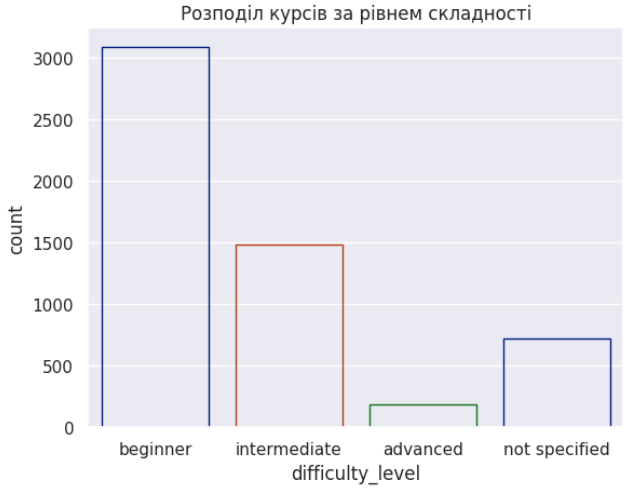

Рисунок 2. Розподіл кількості курсів за рівнем складності

На рис.3 зображено хмару ключових слів та ключових навичок курсів.

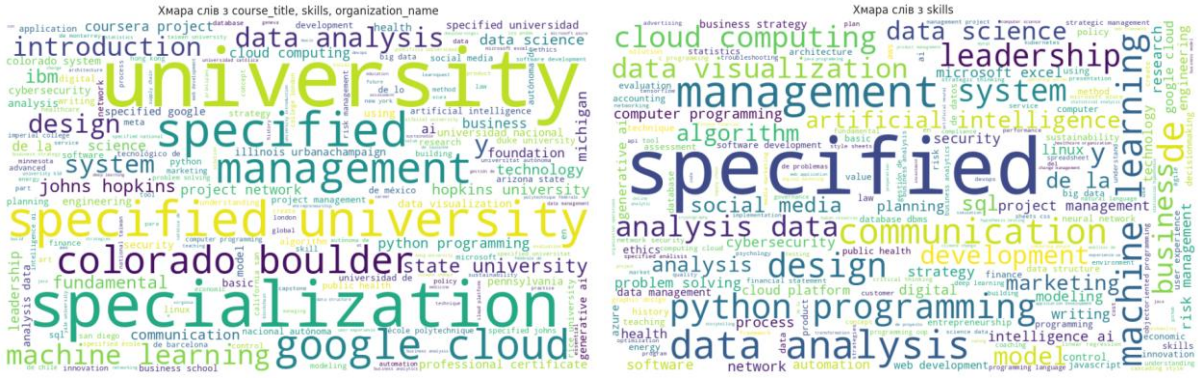

Рисунок 3. Ключові слова та ключові навички курсів

Результати та обговорення

Моделі для прогнозування кількості студентів

Для прогнозування кількості студентів курсу на основі середнього рейтингу, тривалості, кількості модулів, рівня складності та типу розкладу обрано лінійну множинну регресію [3], поліноміальну множинну регресію [4], Random forest [5] та XGBoost [6].

Результати порівняння моделей представлено в таблиці 1. Як бачимо, поліноміальна регресія не покращує якість моделі, R^2 залишається низьким, із зростанням порядку більше 5 навіть зменшується, що свідчить про перенавчання. MAE та RMSE також зростають із підвищенням порядку полінома. Random Forest із глибиною 7 — найкраща модель за всіма метриками, вона має найнижчі MAE та RMSE, а також найвище $R^2 = 0.30$. XGBoost не є кращим за Random Forest, хоча все ще кращий за лінійні підходи. Тому найкращим способом прогнозування в цьому випадку є модель Random forest з максимальною глибиною 7.

Таблиця 1.

Модель	R^2	MAE	MSE	RMSE	Примітка
Множинна лінійна регресія	0.18	10533.68	254265197	15945.7	Було обрано як нижню межу якості
Множинна поліноміальна регресія (порядок 2)	0.18	10844.34	279359914	16714.06	Якість гірша за нижню межу. Модель однозначно не підходить
Множинна поліноміальна регресія (порядок 5)	0.22	10631.42	265673467	16299.49	R^2 дещо вищий за нижню межу, інші показники трохи нижчі. Після 5-го порядку показники погіршуються – починається перенавчання
Random forest (максимальна глибина = 5)	0.28	10079.61	223950086	14964.96	Краще, ніж лінійна регресія
Random forest (максимальна глибина = 7)	0.3	9795.46	216922200	14728.28	Глибина з найкращими показниками
XGBoost	0.26	10097.45	229367392	15144.88	Краще, ніж лінійна регресія, але гірше за Random forest

Порівняємо ефективність цих самих моделей на більш “нормалізованих” даних (вибірка з кількістю студентів до 40000). Середнє значення num_of_students у вибірці – 14283. Обмеження вибірки покращило результати для всіх моделей, збільшило R^2 та зменшило MAE та RMSE. Це доводить, що чим більш однорідні дані і чим менше в них шуму, тим краще працюють моделі прогнозування. Найкращим способом прогнозування в цьому випадку є модель Random forest з максимальною глибиною 8.

Візьмемо другу частину датасету і подивимося, як моделі працюватимуть на таких даних. Середнє значення num_of_students у вибірці – 63664. R^2-метрики для всіх моделей дуже низькі (близькі до 0 або навіть негативні), що свідчить про погану здатність моделей пояснити дисперсію цільової змінної. Поліноміальна регресія 7-го порядку – це єдина модель, що досягає відносно найкращого R^2=0.16, але все одно це дуже поганий результат. Random Forest та XGBoost сильно перенавчаються — результати стають навіть гіршими, ніж у простих моделей. Щоб покращити ситуацію, можна додати більше ознак або розділити датасет на ще кілька менших датасетів.

 ISSN 1560-8956

Моделі для класифікації курсів

Для класифікації курсів на популярні (більше 20000 студентів) та непопулярні (менше 20000 студентів) обрано логістичну регресію [7], Decision tree [8], Random forest [5] та SVM [10].

Серед метрик оцінювання якості precision, recall, f1 score та accuracy в першу чергу будемо звертати увагу на precision. Метрика precision оцінює, який відсоток курсів, віднесених до певного класу справді належить цьому класу. В даній задачі вона є пріоритетною тому, що зазвичай люди обирають саме популярні курси для навчання, організації готові витрачати на рекламу і фінансування популярних курсів більше грошей, тому якщо модель класифікує курс як популярний, а він не стане популярним, це може призвести до втрати бюджету і розчарування клієнтів. Результати порівняння моделей представлено в таблиці 2.

Таблиця 2.

Модель	precision (непоп/поп)	recall (непоп/поп)	f1 score (непоп/поп)	accuracy
Логістична регресія	0.77; 0.64	0.72; 0.7	0.74; 0.67	0.71
Дерево рішень (глибина 5)	0.81; 0.71	0.78; 0.75	0.79; 0.73	0.76
Random forest	0.81; 0.72	0.79; 0.75	0.8; 0.74	0.77
SVM (kernel='linear', C=1)	0.82; 0.64	0.67; 0.79	0.74; 0.71	0.72
SVM (kernel='rbf', C=100)	0.8; 0.66	0.71; 0.76	0.75; 0.7	0.73

Важливим показником якості є precision саме на популярних популярних курсах, оскільки помилкова позитивна класифікація (false positive) може призвести до марного вкладення ресурсів у непопулярні курси. У таблиці можна побачити, що найкращий precision класифікації популярних курсів у моделі Random forest (0.72), а логістична регресія – найгірший (0.64). Також у Random forest високі recall, f1 score та загальна accuracy. Ця модель добре збалансована та мінімізує хибні спрацювання при прогнозуванні популярності.

SVM в обох випадках має високий recall, але precision помітно нижча, а це означає, що багато непопулярних курсів будуть помилково класифіковані як популярні, що суперечить основній меті. Тому ця модель не підходить.

Decision tree має трохи нижчу точність, ніж Random Forest, але простіше інтерпретується.

Логістична регресія є найпростішою, але показує гірші результати.

Також порівняємо confusion matrix усіх методів (рис.4). Відповідно до confusion matrix найменше непопулярних курсів були класифіковані як популярні у Random forest, що підтверджує зроблений вище висновок.

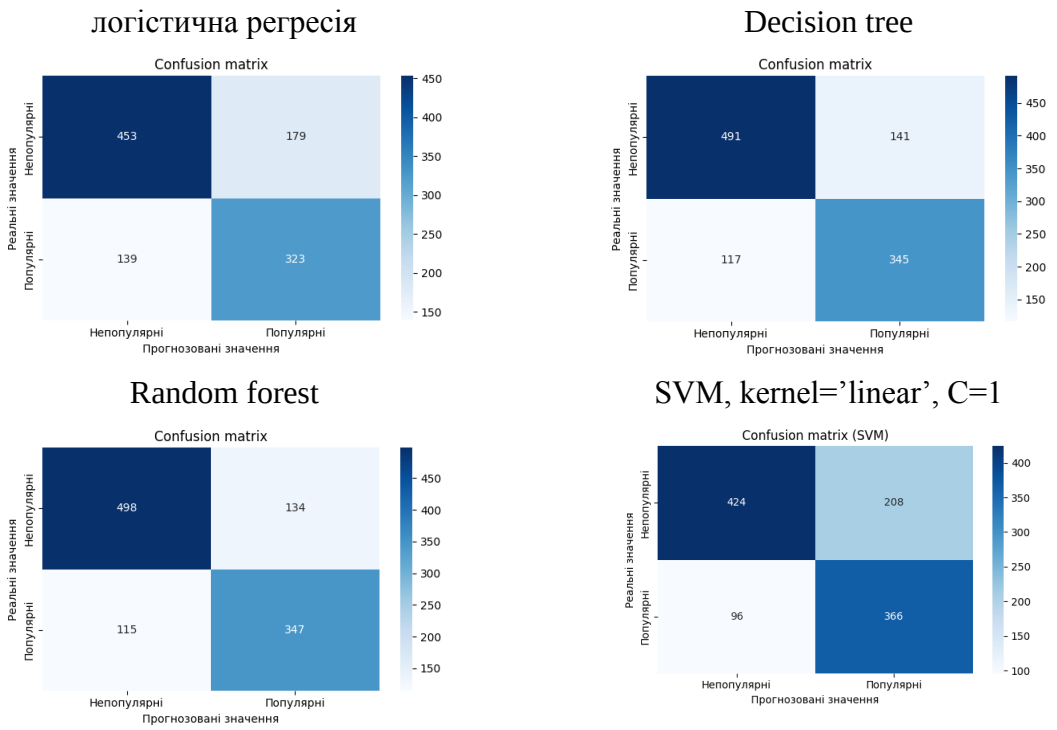

Рисунок 4. Матриці невідповідностей для моделей класифікації

Як і в попередньому дослідженні, порівняємо моделі для датафрейму з курсами, де кількість студентів не перевищує 40000. В цьому випадку Random forest теж показала найвищу precision для популярних курсів (0.64), хороший баланс між precision і recall і найвищу accuracy. Decision tree також непогана модель. SVM в обох випадках поступається Random forest. Логістична регресія показала найнижчий результат і може використовуватися лише як нижня межа якості для порівняння. Відповідно до confusion matrix найменше непопулярних курсів були класифіковані як популярні у Random forest, що підтверджує зроблений вище висновок. Логістична регресія теж менше помилилася щодо класифікації непопулярних як популярних курсів, але вона класифікувала багато популярних курсів як непопулярні, що теж може призвести до помилок.

Висновки

Дослідивши поведінку моделей прогнозування на різних вибірках, можна сказати, що найкращий баланс показників R^2, MAE, RMSE та MSE у нормалізованій вибірці (кількість студентів <= 40000). У цій вибірці дані мають меншу дисперсію, тому моделі можуть ефективно захопити закономірності. Поліноміальні та ансамблеві моделі показують помітно кращу якість. Цей датафрейм можна використовувати як основну підвибірку для розробки надійної моделі. Для повного датасету моделі показують середню якість. Вони частково розпізнають тренди, але розкид даних великий. "Хвости" в розподілі ускладнюють точне прогнозування. Датасет можна використовувати для загальної оцінки моделі, але високу точність не гарантує. На великій вибірці показники якості дуже погані, ансамблеві моделі сильно перенавчаються та не здатні

ISSN 1560-8956

адекватно узагальнювати. Цільову змінну неможливо добре пояснити з доступними даними, можливо, варто розширити набір ознак.

В обох розглянутих випадках класифікації курсів на популярні та непопулярні, найкращі результати стабільно демонструє модель Random Forest. Вона забезпечує найвищий precision для популярних курсів — ключової метрики для даної задачі, адже помилкова класифікація непопулярного курсу як популярного може призвести до фінансових втрат. Random Forest також демонструє збалансовані значення recall, f1-міри та загальної точності accuracy як на повному датасеті, так і на датасеті із кількістю студентів до 40000. Дерево рішень також є простою й інтерпретованою моделлю з непоганою якістю, але поступається Random Forest. Моделі SVM і логістична регресія мають нижчі показники precision для класу "популярний" і менш стабільну ефективність. Таким чином, Random Forest є найнадійнішою моделлю в рамках цієї задачі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Офіційний сайт Coursera. URL: https://www.coursera.org/programs/program-natsional-nii-tiekhnichnii-univiersitiet-ukrayini-kiyivs-kii (дата звернення: 21.05.2025)

2. Вебресурс Kaggle. URL: https://www.kaggle.com/datasets (дата звернення: 21.05.2025)

3. *Seber, G. A. F., Lee, A. J.* Linear Regression Analysis [Електронний ресурс]. – 2-ге вид. Wiley, 2003. 512 ст.

4. *Al-Kasasbeh, M., Al-Azzam, N., Al-Momani, A.* Modeling with polynomial regression [Електронний ресурс] // Procedia Computer Science 2012. Т. 65. С. 426–432. URL: https://www.sciencedirect.com/science/article/pii/S1877705812046085 (дата звернення: 30.05.2025)

5. *Biau, G., Scornet, E.* A random forest guided tour [Електронний ресурс] // TEST. 2016. Т. 25. С. 197–227. URL: https://link.springer.com/article/10.1007/s11749-016-0481-7 (дата звернення: 30.05.2025)

6. *Chen, T., Guestrin, C.* XGBoost: A scalable tree boosting system [Електронний ресурс] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York: ACM, 2016. С. 785–794. URL: https://dl.acm.org/doi/abs/10.1145/2939672.2939785 (дата звернення: 30.05.2025)

7. *LaValley, M.P.* Logistic regression [Електронний ресурс] // Circulation. – 2008. Т. 117, № 18. С. 2395–2399. URL: https://doi.org/10.1161/CIRCULATIONAHA.106.682658 (дата звернення: 30.05.2025)

8. *Song, Y.-Y., Lu, Y.* Decision tree methods: applications for classification and prediction [Електронний ресурс] // Shanghai Arch Psychiatry. – 2015. Т. 27, № 2. С. 130–135. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/ (дата звернення: 30.05.2025)

9. *Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. та ін.* Support Vector Machines [Електронний ресурс] // Scikit-learn: Machine Learning in Python. URL: https://scikit-learn.org/stable/modules/svm.html (дата звернення: 30.05.2025)