ФІЗИКО-МАТЕМАТИЧНІ НАУКИ

Бережняк М.О.
студент,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

ПРОГНОЗУВАННЯ УСПІШНОСТІ СТУДЕНТІВ ОНЛАЙН-КУРСІВ

Prometheus [1] – це сервіс масових відкритих онлайн курсів, побудований на платформі Open edX, що розроблена Масачусетським інститутом технологій спеціально для створення МВОК. У вересні 2014 року на платформі Prometheus з'явилися 2 перших масових відкритих онлайн-курси українською мовою: «Фінансовий менеджмент» [2] та «Історія України: від Другої світової війни до сучасності» [3].

Мета будь-якої навчальної онлайн-платформи – збільшувати освіченість населення. Світова практика показує, що серед усіх зареєстрованих на онлайн-курс користувачів, в середньому його завершують лише 10% студентів. Виникає природня потреба у заохоченні студентів, підвищення їхньої мотивації під час проходження курсу. Проблема є комплексною, оскільки спочатку треба ідентифіковувати неуспішних користувачів, а потім вживати заходів, щоб їх мотивувати на завершення курсу. Ці заходи можуть варіюватися від звичайної розсилки по електронній пошті нагадувань до персоналізованих змін у інтерфейсі онлайн-платформи, структурі курсу, його змісті тощо [4]. Тут буде розглянуто підходи до прогнозування успішності користувачів. Як вихідні дані буде використаймо інформацію про зазначені вище курси історії та фінансового менеджменту. Всі дані добровільно керівництвом Prometheus.

Користувачі платформи Prometheus навчаються за вільним графіком: під час старту курсу навчальні матеріали надходять щотижнево, доки не досягнуть кінця курсу, потім вони лишаються відкритими для всіх зареєстровиних на курс користувачів. Обмежень часових по проходженню курсу немає.

У системі Prometheus користувачі навчаються, і для них характерна певна поведінка і результати успішності. Всього система має певну кількість користувачів користувачів. Результати успішності у правильності відповідей на запитання тестів виражаються у оцінці за кожний тест. Якщо користувач не зробив спроби проходження тесту – оцінка рівна 0. За проходження курсу студенти отримають сертифікат. Відповідно всі студенти курсу поділяються на тих, хто має сертифікат, і тих, хто не має. Для успішних і неуспішних студентів характерні певні типи поведінки і результативність. Надійною характеристикою системи можна вважати оцінки проходження тестів. Варто зазначити, що у кожного користувача може бути різна кількість оцінок, бо студенти не забов'язані проходити курс синхронно.

Таким чином проблема полягає у прогнозі успішності студента, на основі наявних оцінок у певний момент часу. Маємо задачу бінарної класифікації [5]: треба визначити завершить студент курс чи ні. Проведемо формалізацію задачі. Дано:

– кількість користувачів – s;
– кількість тестів для курсу – m;
– наявність сертифікату у *i*-го користувача – $c^i, i = \overline{1,s}$;
– оцінка проходження *j*-го тесту *i*-м користувачем – $g_j^i, j = \overline{1,m}, i = \overline{1,s}$.

Вихідні дані: булева(1 – так, 0 – ні) оцінка проходження курсу користувачем – $\hat{c}$. Об'єктом у даній задачі є користувач. X – простір характеристик об'єкта.

$$X = \{ g_j^i, j = \overline{1,m}, i = \overline{1,s} \} \quad (1)$$

Y = {1,0} – множина класів(пройшов/не пройшов курс).

$$y^* : X \to Y \quad (2)$$

(2) – цільова залежність, що виражається через $c^i, i = \overline{1,s}$. Задача: знайти залежність (3), де $x \in X$, $y \in Y$; знайти $f(x) = \hat{c}$.

$$f : X \to Y \quad (3)$$

Проведемо дослідження кращого алгоритму на вибірках даних з курсів «Фінансовий менеджмент» та «Історія України: від Другої світової війни до сучасності». Як мову програмування оберемо Python, оскільки для цієї мови розроблені зручні бібліотеки для наукових розрахунків та машинного навчання[6]. Експерименти будемо проводити у середовищі IPython Notebook. Використаймо реалізаці алгоритмів з бібліотеки scikit-learn. Для оцінки точності використаємо кросс-валідацію[7]. За змінні для прогнозування оберемо результати перших чотирьох тестів. Враховуючи те, що в загальносвітовій практиці онлайн курс завершуються в середньому 10% студентів, будемо порівнювати побудовані моделі з контрольною, яка завжди каже, що користувач курс не пройде. Результати експерименту наведено у таблиці 1:

Таблиця 1

Дослідження точності роботи алгоритмів

Алгоритм	**Курс історії. Точність моделі**	**Курс фінансового менеджменту. Точність моделі**
Логістична регресія	0.954	0.956
Наївний баєсівський класифікатор	0.956	0.954
K найближчих сусідів	0.944	0.952
Дерева рішень	0.949	0.951
Метод опорних векторів	0.955	0.955
Контрольна модель	0.924	0.894

У випадку курсу фінансового менеджменту найкращий результат показала логістична регресія, у випадку історії – наївний баєсівський класифікатор. Контрольна модель показує точність 0.89 для фінансового менеджменту та 0.92 для історії. Це означає, що курс пройшли 11% і 8% користувачів відповідно. Всі алгоритми показали точність порядку 0.95 для обох курсів. Проте різниця серед алгоритмів виявляється лише на тисячний долях точності, а це означає їхню практичну еквівалентність при розв'язанні поставленої задачі. Для визначеності обремо логістичну регресію як основний алгоритм.

Для попереднього дослідження ми використовували результати перших чотирьох пройдених тестів. Проте на практиці структура курсу може відрізнятися: тестів може бути більше або менше. Проведемо дослідження залежності точності моделі від кількості ознак, на якій вона побудована. Контрольну модель використаємо ту саму. Повні результати експерименту наведені у таблиці 2. Для наочності побудуємо діаграми – рисунок 1 і рисунок 2:

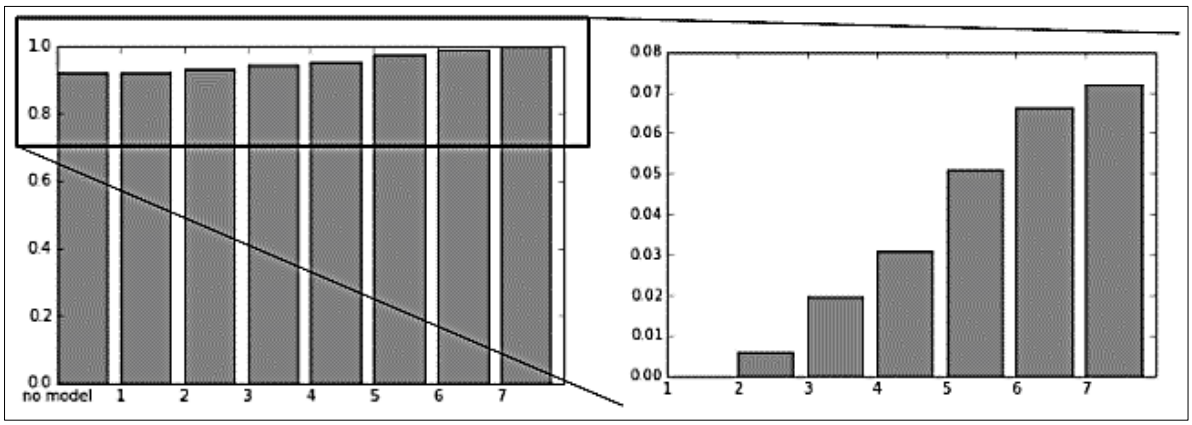

Рис. 1. Курс історії. Якість прогнозу

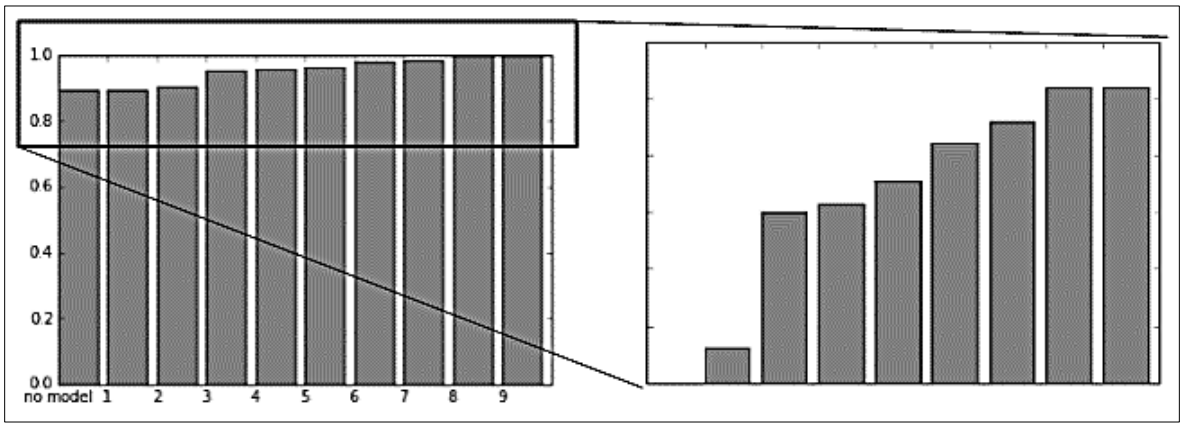

Рис. 2. Курс фінансового менеджменту. Якість прогнозу

На рисунку 3 та рисунку 4 зображено точність прогнозу в залежності від кількості атрибутів моделі. Очевидно, що чим більше атрибутів, тим більше точність. Для курсу історії точність росте монотонно, а для курсу фінансового менеджменту стрибкоподібно(за допомогою результатів перших трьох тестів можна досягти хорошої точності прогнозу). Отже, точність в залежності від

кількості атрибутів може рости по-різному для різних курсів. Результати експерименту наведено у таблиці 2.

Таблиця 2

Дослідження точності роботи алгоритму для різної кількості атрибутів

Кількість атрибутів	Курс історії. Точність моделі	Курс фінансового менеджменту. Точність моделі
Контрольна модель	0.924	0.893
1	0.924	0.893
2	0.930	0.905
3	0.943	0.953
4	0.954	0.956
5	0.974	0.964
6	0.990	0.978
7	0.996	0.985
8	-	0.997
9	-	0.997

Отже, маємо задачу прогнозування успішності студентів онлайн-курсів в залежності від кількості оцінок, які є наявні. Їх може бути довільна кількість. Поставлена задача зводиться до задачі бінарної класифікації, що може бути розв'язана багатьма методами, наприклад, логістичною регресією, яка стабільно добре працює при різних наборах даних. Чим більше оцінок студента відомо, тим більша точність прогнозу успішності. Варто відмітити, що прогнозування також має сенс при невеликій кількості відомих оцінок. Це доводить випадок курсу фінансового менеджменту, де вже при відомих результатах трьох тестів(із дев'яти можливих) можна з точністю 95% говорити, пройде студент курс чи ні. Такі результати залежать від конкретного набору даних конкретного курсу, і їх важко передбачити.

Список використаних джерел:

1. Прес-реліз проекту Prometheus [Електронний ресурс]. – 2014. – Режим доступу до ресурсу: http://prometheus.org.ua/prometheus-start/

2. Курс фінансового менеджменту [Електронний ресурс] – Режим доступу до ресурсу: http://courses.prometheus.org.ua/courses/NAUKMA/101/2014_T2/about

3. Курс історії [Електронний ресурс] – Режим доступу до ресурсу: http://courses.prometheus.org.ua/courses/KNU/101/2014_T2/about

4. Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. IEEE Transactions on Knowledge and Data Engineering / Perera, Kay, Koprinska та ін.], 2009. – (21). – С. 759-772.

5. Statistical classification [Електронний ресурс] – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Statistical_classification

6. Andrew N. Machine Learning [Електронний ресурс] / N. Andrew – Режим доступу до ресурсу: https://www.coursera.org/learn/machine-learning.

7. Schutt R. Doing Data Science / R. Schutt, C. O'Neil., 2014.