6. Garcia M., Rodriguez A. Mobile learning and its impact on educational accessibility // Mobile Education. – 2017. – Vol. 22(3). –Pp. 123–138.

7. Smith J. The impact of interface design on educational effectiveness // Journal of Educational Technology. – 2018. – Vol. 45(2). – Pp. 167–183.

УДК 519.2+51-77
doi:10.18720/SPBPU/2/id24-45

Aleksandra D. Danilova [1],
Bachelor;
Vadim G. Pak [2],
Associate Professor, Candidate of Physical and Mathematical Sciences

METHODS OF DATA ANALYSIS IN THE PROBLEM OF RETAINING AND ATTRACTING THE AUDIENCE OF THE MASSIVE OPEN ONLINE COURSE

[1, 2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia; [1] danilova.ad@edu.spbstu.ru; [2] pak_vg@spbstu.ru

Abstract. The paper investigates the results of passing a massive open online course "Methods of Computational Mathematics" to retain and attract an audience. It is shown that the completion rate differs for different categories of students, namely those who independently enrolled in the course and took the course within the framework of the compulsory academic discipline "Educational Foresight". The simplest and most difficult tasks common to both categories of students have been identified. Recommendations on course modification are offered.

Keywords: data analysis, massive open online course, audience retention, retention metric, histogram, distribution, statistical methods of data analysis.

Данилова Александра Дмитриевна [1],
бакалавр;
Пак Вадим Геннадьевич [2],
доцент, доцент, канд. физ.-мат. наук

МЕТОДЫ АНАЛИЗА ДАННЫХ В ЗАДАЧЕ УДЕРЖАНИЯ И ПРИВЛЕЧЕНИЯ АУДИТОРИИ МАССОВОГО ОТКРЫТОГО ОНЛАЙН КУРСА

[1, 2] Россия, Санкт-Петербург, Санкт-Петербургский политехнический
университет Петра Великого,
[1] danilova.ad@edu.spbstu.ru; [2] pak_vg@spbstu.ru

Аннотация. В статье исследуются результаты прохождения массового открытого онлайн-курса «Методы вычислительной математики» с целью удержания и привлечения аудитории. Показано, что показатели успеваемости различаются у разных категорий студентов, а именно у тех, кто самостоятельно записался на курс и прошел курс в рамках обязательной учебной дисциплины «Образовательный форсайт». Были определены самые простые и самые сложные задания, общие для обеих категорий студентов. Предлагаются рекомендации по модификации курса.

Ключевые слова: анализ данных, массовый открытый онлайн-курс, удержание аудитории, показатель удержания, гистограмма, распределение, статистические методы анализа данных.

Introduction

Every year in the world, including in Russia, the number of massive open online courses (MOOCs) increases. Thanks to their wide reach, they help expand access to quality education, attract new students and strengthen the brand of the universities [1, 7].

The purpose of this work is to study the results of passing the MOOC "Methods of Computational Mathematics" from the Higher School of Artificial Intelligence Technologies of Peter the Great St. Petersburg Polytechnic University on the Open Education platform to retain and attract an audience. In accordance with the goal, the following work tasks are defined:

A. To study the results of completing the course over several sessions.

B. To offer recommendations on course modification.

1. Course description

Let's take a closer look at the structure of the course "Methods of Computational Mathematics".

There are two categories of students on the course:

– students taking the course within the framework of the compulsory academic discipline "Educational Foresight";

– students who self-registered for the course.

The course consists of 15 topics divided into 4 modules. The course includes the following assignments:

– 15 current control tests;

– 9 laboratory works;

– 4 midterm tests for each module.

The course is considered successfully completed if more than 50 % of the points are scored in total.

The results of the sessions were collected for the period from fall 2016 to spring 2023. Session reports contain only student IDs and the points they scored for each assignment.

2. Study of the results of the course

2.1. Retention metric

Retention metric *R*, or completion rate is the percentage of people who successfully completed the course (in accordance with the standards specified by the teacher) from among those who were enrolled in it [5].

Note that this metric does not reflect the degree of usefulness of the course, as well as the variety of goals and models of student engagement [5], so we will calculate it separately for each category of students. The results for students who enrolled in the course by themselves (1) and students who were taking part in the "Educational Foresight" (2) are given below.

$$R_{\text{self-registered}} = 2.8\ \% \tag{1}$$

$$R_{\text{foresight}} = 74.1\ \% \qquad (2)$$

Let’s investigate the assignments with the aim of modifying them to increase the metric in both cases.

2.2. Course assignments

Figures 1 and 2 show histograms of average grades on assignments for both categories of students: the first histogram shows average scores on current tests, the second histogram — on laboratory work, the third histogram — on midterm tests.

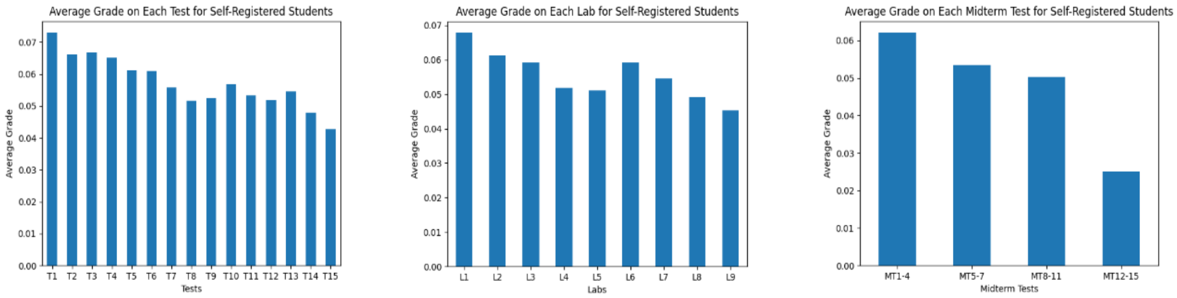

Fig. 1. Average grades on course materials for students who enrolled on the course by themselves

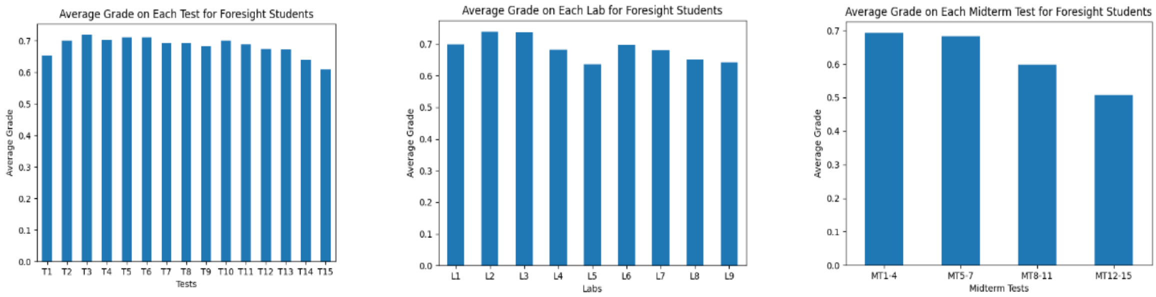

Fig. 2. Average grades on course materials for students who took the course during “Educational Foresight”

The histograms show “waves” when academic performance is high at first, then begins to decline. Let’s select two “waves” on the histograms for their more detailed study, so that in the future they can be aligned according to the proposed modifications of assignments in such a way that the course becomes more accessible to the prepared students and it can be passed more evenly and smoothly. Let’s denote wave 1 (distribution 1) in green, and wave 2 (distribution 2) in red as shown in Figure 3.

Let’s compare the forms of distributions between categories.

Let’s use the Kolmogorov-Smirnov test to test the hypothesis that the distribution functions are similar. The distance [2] between the empirical distribution functions $H_1(x)$ and $H_2(x)$ is calculated as follows in the formula (3) below:

$$D = \sup_x |H_1(x) - H_2(x)|. \qquad (3)$$

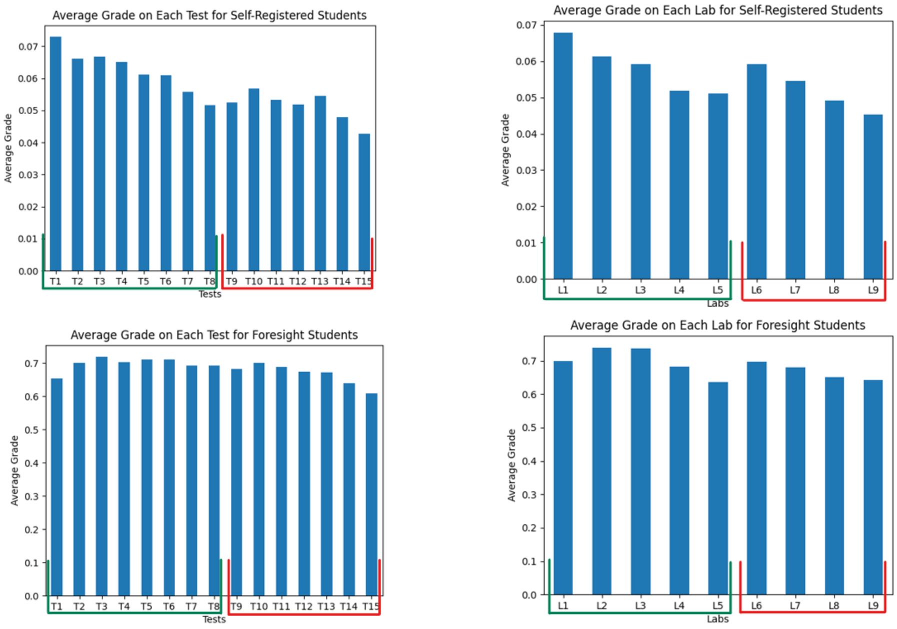

Fig. 3. Waves (distributions)

According to Table 1, the p-value is less than the threshold of 0.05. This means that the hypothesis of similarity of the distribution forms is rejected.

Table 1

Kolmogorov-Smirnov test for comparing pairs of distributions

Assignment Type	p-value for Distributions № 1	p-value for Distributions № 2
Current tests	0.0002	0.0006
Laboratory works	0.0079	0.0286

The Bhattacharyya distance [3, 6] allows to measure the similarity of two probability distributions *P* and *Q*, and is calculated by the formula (4):

$$D = -\ln\left(\sum_i \sqrt{P(i)Q(i)}\right). \tag{4}$$

It takes values from the interval [0; +∞), where 0 is responsible for the total similarity of distributions.

The Bhattacharyya distance was calculated for two types of distributions between student categories. It confirms the result obtained by testing the Kolmogorov-Smirnov test, and it says the same thing, namely the forms for both categories of students are not identical. The obtained results are present in Table 2.

Table 2

Bhattacharyya distance test for comparing pairs of distributions

Assignment Type	Distributions № 1	Distributions № 2
Current tests	1.09	1.51
Laboratory works	0.63	0.45

Let’s find the kurtosis and skewness in order to draw conclusions about the nature of the differences in the transition from test to test and from laboratory work to laboratory work. This will allow to see later how seriously the histograms will have to be aligned when comparing two categories of students, for a smooth passage of the course in prospect.

The kurtosis coefficient characterizes the degree of peaking of the distribution [4], the skewness coefficient characterizes the measure of the asymmetry of distribution shape to the left or right relative to the symmetrical one [2]. A negative value of the kurtosis coefficient indicates a smaller number of outliers compared to a normal distribution.

According to the results obtained in Tables 3 and 4, for current tests the shape of the distributions in both categories of students is almost the same, there are no outliers; the distributions are skewed, which confirms the unevenness of the distributions.

Table 3

The study of the similarity of distributions on current tests for both categories of students

Student Category	Measure	Distributions № 1	Distributions № 2
Self-registered	Kurtosis	-1.50	-1.50
	Skewness	-0.19	0.69
Foresight	Kurtosis	-1.50	-1.50
	Skewness	-0.42	0.69

Table 4

The study of the similarity of the distributions of laboratory works for both categories of students

Student Category	Measure	Distributions № 1	Distributions № 2
Self-registered	Kurtosis	-1.50	-1.49
	Skewness	-0.40	0.53
Foresight	Kurtosis	-1.49	-1.50
	Skewness	0.28	-0.71

Let’s identify assignments common to both categories of students in each wave that require modifications in order to align the histogram (but in such a way that complex assignments are made more accessible to a prepared, trained student, and make easy ones more difficult).

Table 5 shows the common peaks and declines for both categories of students (previously the data were sorted). Common declines include tests 15 and 14 (wave 2), laboratory works 5 (wave 1), 9 and 8 (wave 2); common peaks include test 3 and laboratory works 2-3 (wave 1).

Table 5

Declines and peaks common to both categories of students

Student Category	Current Tests		Laboratory Works	
	Top 3 (min)	Top 3 (max)	Top 3 (min)	Top 3 (max)
Self-registered	T15 0.042 T14 0.047 T8 0.051	T1 0.072 T3 0.066 T2 0.066	L9 0.045 L8 0.049 L5 0.051	L2 0.061 L3 0.059 L6 0.059
Foresight	T15 0.608 T14 0.638 T1 0.652	T3 0.718 T5 0.711 T6 0.710	L5 0.636 L9 0.642 L8 0.651	L2 0.738 L3 0.737 L1 0.700

The Kolmogorov-Smirnov test confirms the significance of the coincidences of the top (both common and uncommon) peaks and declines (since the plots are monotonic linear functions, the p-value equals to 0.09). Consequently, there is correlation between the peaks / declines common to both groups of students, which is confirmed by Spearman’s correlation (for tests and laboratory works № 2-3, it is equal to 1).

2.3. Recommendations for course modification

Among the previously identified assignments, the recommendations for modifying assignments were highlighted. For example, the use of additional interpolation methods, reducing the dimensionality of problems and requirements for the accuracy of solutions. A detailed description is given below.

Complications:

– test 3 (topic “Approximation. The least squares method”): add questions and tasks on non-polynomial interpolation (for example, choosing the best model from a given list: exponential, logarithmic, fractional-rational, etc.);

– midterm test 1-4: the same as stated above;

– laboratory work 2 (topic “Splines”): add non-bicubic spline problems with the derivation of equations for finding coefficients;

– laboratory work 3 (topic “Trigonometric interpolation”): add different interpolation methods and other fast Fourier transform algorithms (besides thinning).

Simplifications:

– test 15 and laboratory work 9 (topic “Solving boundary value problems for ordinary differential equations”): add time to take the test, reduce the volume of calculations due to a smaller grid, reduce the requirements for the accuracy of the solution;

– test 14 and laboratory work 8 (topic “Methods for solving differential equations and systems”): reduce the order or dimension of the equations;

– midterm test 12-15: the same as stated above;

– laboratory work 5 (topic “Numerical methods for solving systems of linear equations”): reduce the requirements for the accuracy of the solution, reduce the dimension of the system (3x3 maximum).

Conclusion

The paper examined the results of completing an online course over several years and proposed changes to the course tasks that could potentially increase the retention metric and contribute to attract an audience.

A methodology for applying statistical data analysis to improve the fitness of the course to a mass audience is proposed and tested.

As further research, it is proposed to implement changes in the course to check how much the retention metric has changed.

References

1. Вилкова К.А., Захарова У.С., Семенова Т.В. Нельзя просто переложить в онлайн традиционные формы обучения // Круглый стол «Онлайн-обучение в вузе: риски и возможности». – 2020. – URL: https://ioe.hse.ru/news/341458289.html (дата обращения: 11.10.2023).

2. Dodge Y. Kolmogorov–Smirnov Test // The Concise Encyclopedia of Statistics. – New York, NY: Springer New York, 2008. – Pp. 283–287.

3. Fukunaga K. Feature extraction and linear mapping for classification // Introduction to Statistical Pattern Recognition (Second Edition). – Chap. 10. – Boston: Academic Press, 1990. – Pp. 441–507.

4. Lee J. Statistics, descriptive // R. Kitchin, N. Thrift (eds.) International Encyclopedia of Human Geography. – Oxford: Elsevier, 2009. – Pp. 422–428.

5. Koller D., Ng A., Do T., Chen Z. Retention and intention in massive open online courses: in depth // EDUCAUSE Review. – 2013. – Vol. 48, No 3. – Pp. 62–63.

6. Safjan K. Understanding Bhattacharyya distance and coefficient for probability distributions [Electronic Source]. – URL: https://safjan.com/understanding-bhattacharyya-distance-and-coefficient-for-probability-distributions/ (date of access: 04.10.2023).

7. Shah D. A decade of MOOCs: a review of MOOC stats and trends in 2021 [Electronic Source]. – URL: https://www.classcentral.com/report/moocs-stats-and-trends-2021/ (date of access: 11.10.2023).