Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Association Analysis of Online Learning Behavioral Characteristics and Academic Achievement

Yunyi Wang

2941308283@qq.com

Fuzhou Melbourne Polytechnic

Article

Keywords: Online learning behavior, learning duration, course completion rate, score distribution, learning analysis, educational data mining

Posted Date: October 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10980696/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Association Analysis of Online Learning Behavioral Characteristics and Academic Achievement

Yunyi Wang [1] *

[1] Department of International Business, Fuzhou Melbourne Polytechnic, Fuzhou, 350108, Fujian, China

***Corresponding author, Email: [2941308283@qq.com]**

June 2026

ABSTRACT

Background: The rapid advancement of educational informatization has positioned online learning as a mainstream educational modality. Nevertheless, persistent challenges — particularly low course completion rates and high learner attrition — continue to constrain its effectiveness.

Objective: This study systematically examines online learning behaviors across three core dimensions — learning duration, course completion rate, and grade distribution — with the aim of identifying actionable strategies to enhance platform performance.

Methods: A mixed-methods design was adopted, integrating literature review, empirical analysis, and case studies. A theoretical framework was constructed, and a behavioral feature set encompassing learning duration, completion rates, grade distribution, and interaction metrics was established. On this basis, predictive models were developed and benchmarked using classification algorithms, ensemble learning, deep learning, and clustering techniques.

Results: Findings indicate a statistically significant positive correlation among learning duration, course completion rates, and grade distribution, confirming their roles as key indicators of online learning effectiveness. From a methodological standpoint, behavior analysis has progressed toward advanced deep learning architectures, with reported prediction accuracies surpassing 95% on benchmark datasets. Distinct behavioral patterns were identified among learners — categorized as high-engagement, low-engagement, skip-pattern, and strategic types. Additionally, instructional design was found to exert a notable influence on behavioral outcomes.

Conclusion: Drawing on these findings, optimization strategies are proposed at three levels: (1) Platform level — strengthen data analytics infrastructure to support early-warning systems and personalized recommendation mechanisms; (2) Instructional design level — prioritize scenario-based learning, resource diversification, interaction enhancement, and feedback optimization; (3) Management level — reinforce behavioral monitoring, implement early-intervention protocols, facilitate instructor training, and establish incentive mechanisms. This study provides an evidence-based foundation for personalized recommendation and curriculum design, offering practical implications for advancing the quality of online education.

Key words: Online learning behavior, learning duration, course completion rate, score distribution, learning analysis, educational data mining

Table

1. Introduction 1

2. Literature review 2

2.1 The theoretical basis and data source of online learning behavior analysis 2
2.2 Characteristics and analysis methods of learning behavior 2
2.3 Relevant research on learning effect prediction 3

3. Theoretical framework 4

3.1 The concept of theory 4
3.1.1 The behavior of online learning 4
3.1.2 Development of learning analysis and educational data 4
3.2 The basis of relevant theories 4
3.2.1 Learn the theory of analytical technology 4
3.2.2 Educational Data Mining Theory 5

4. Research method 6

4.1 Data Collection and Processing 6
4.2 Behavioral Characteristics 6
4.3 Predictive Model Construction 7
4.3.1 Classification Algorithm 7
4.3.2 Ensemble Learning Algorithm 7
4.3.3 Deep Learning Algorithm 7
4.4 Cluster Analysis Method 7

5. Case analysis 8

5.1 Descriptive statistical analysis of learning behavior 8
5.1.1 The relationship between study time and grades 8
5.1.2 Analysis of the characteristics of major online learning platforms 10
5.2 Analysis of learning time and learning frequency 12
5.2.1 Distribution of learning time 12
5.2.2 Distribution of learning frequency 14
5.3 Course completion rate and score distribution analysis 15
5.4 Analysis of the market size of online education 16
5.5 Evidence of the learning effect based on the prediction model 18
5.6 Cluster analysis of learner group characteristics 19
5.7 Online teaching design and behavioral intervention revelation 19

6. Conclusion and Discussion 21

6.1 Summary of Findings 21
6.2 Implications for Practice 21

6.3 Limitations and Future Work 22

Declarations 23

Special Reminder 23
Competing interests 23
Ethics approval and consent to participate 23
Availability of data and materials 23
Author contributions 24
Funding 24

7. References 25

1. Introduction

The rapid proliferation of computer technologies, the Internet, and digital media has established online learning as a transformative force in the educational landscape, fundamentally reshaping traditional pedagogical paradigms[1]. The acceleration of educational informatization has given rise to diverse technology-enhanced teaching methodologies, culminating in the generation of vast educational big data[7]. While online learning offers learners unprecedented flexibility and convenience, it is concurrently plagued by persistent challenges, most notably low course completion rates and high student attrition[8].

Consequently, the analysis of extensive learning behavior data to uncover underlying learning patterns holds significant theoretical and practical value for enhancing the quality of online education and optimizing instructional design. By leveraging educational data mining and learning analytics, educators and platform developers can gain deeper insights into the learning process, moving beyond simple content delivery to a more data-driven approach.

This study focuses on the systematic analysis of online learning behavior, centering on three core dimensions: learning duration, course completion rate, and score distribution. The primary objective is to evaluate the effectiveness of online learning platforms and explore avenues for their optimization. Through an in-depth investigation of learners' behavioral characteristics, this research aims to provide a scientific basis for personalized learning recommendations and evidence-based curriculum design, thereby offering actionable guidance for teaching practices and platform management.

2. Literature review

2.1 The theoretical basis and data source of online learning behavior analysis

Online learning behavior analysis is fundamentally anchored in educational big data and learning analytics technologies. The ongoing process of educational informatization has generated a vast corpus of data, elevating the role of data from a passive record to a strategic asset[7]. This data, often structured within complex network information systems, provides a granular reflection of student learning behaviors, which in turn directly influence learning outcomes[1].

Data sources for such analyses predominantly encompass a variety of online teaching platforms, including Massive Open Online Course (MOOC) platforms like Chaoxing (Superstar Learning). These platforms meticulously log user interactions, such as video completion rates and task engagement. For instance, Gao and Ding utilized[6] Python-based web crawling techniques to harvest student online learning data, establishing a robust empirical foundation for behavior analysis. Similarly, Wang et al. conducted[4] an in-depth analysis of student behavioral patterns within the Chaoxing platform ecosystem, focusing on the course teaching process.

2.2 Characteristics and analysis methods of learning behavior

The methodological landscape for mining learning behavioral characteristics has undergone a significant evolution, transitioning from traditional statistical analysis to advanced machine learning and deep learning paradigms.

Statistical Analysis: Early research primarily relied on descriptive statistical methods. For example, Lu et al. employed content analysis on the online learning logs of 193 learners, utilizing ROST CM 6.0 for word frequency analysis and semantic network mapping to identify behavioral trends.[5]

Machine Learning: With technological advancements, machine learning methods were progressively integrated. Dai et al. applied[9] ensemble learning techniques to predict academic

performance. Furthermore, Wu enhanced the Stacking algorithm to construct a more robust classification model, demonstrating improved predictive capabilities.[8]

Deep Learning: In recent years, deep learning methodologies have gained prominence. An et al. proposed[3] a CNN-LSTM prediction model incorporating an attention mechanism. In their approach, student learning behavior data was normalized and processed via correlation analysis before being fed into the model. The attention mechanism effectively weighted the importance of different information inputs, thereby significantly enhancing the model's accuracy and reliability.[3]

2.3 Relevant research on learning effect prediction

Predicting learning outcomes represents a critical application of online learning behavior analysis.

Early Approaches: Initial studies widely adopted multivariate regression algorithms. Gao and Ding, for instance[6], performed multidimensional calculations on student learning behavior data to predict learning efficacy.

Current Mainstream: Currently, ensemble learning and deep learning have emerged as the dominant methodologies. Wu utilized an improved[8] Stacking algorithm to complete classification and prediction tasks on the edX public dataset, achieving improvements of 2.46%, 2.11%, and 1.97% in accuracy, precision, and F1-score, respectively. Corroborating this trend, An et al. proposed a performance prediction model that achieved a prediction accuracy of 95.69%[3]. Additionally, Meng et al. employed classifiers to estimate learning outcomes, providing students with actionable feedback on their learning status[1].

3. Theoretical framework

3.1 The concept of theory

3.1.1 The behavior of online learning

Online learning behavior refers to the observable and recordable activities undertaken by learners within a digital environment to achieve specific learning objectives. These activities encompass a wide range of interactions, including but not limited to course access, video consumption, submission of online assignments and quizzes, and participation in discussion forums. This behavioral data serves as a digital footprint, providing an authentic reflection of the learner's engagement level, investment, and the specific learning strategies they employ throughout the educational process[8].

3.1.2 Development of learning analysis and educational data

Learning Analytics (LA) involves the measurement, collection, analysis, and reporting of data about learners and their contexts for the purpose of understanding and optimizing learning and the environments in which it occurs. Educational Data Mining (EDM), while closely related, focuses more on developing methods for exploring the unique and increasingly large-scale data that come from educational settings. The primary distinction lies in their focus: LA emphasizes real-time feedback and intervention to improve learning, whereas EDM is more concerned with model discovery and predictive analytics[10].

3.2 The basis of relevant theories

3.2.1 Learn the theory of analytical technology

The theory of Learning Analytics Technology underscores the use of technical means to standardize and enhance the teaching process. It posits that by collecting granular, real-time data on student activities, educators can move beyond retrospective analysis to provide targeted,

benign interventions that improve learning effectiveness[7]. This theoretical perspective champions the idea of "activating dormant data" through big data technologies. In this paradigm, the teacher's role evolves from a mere lecturer to a mentor or facilitator—particularly in models like the flipped classroom—guiding students to become more effective and self-regulated learners.

3.2.2 Educational Data Mining Theory

Educational Data Mining theory is grounded in machine learning, statistics, and data mining methodologies. Its core tenet is to uncover hidden patterns, structures, and laws within educational data through in-depth exploration. This process emphasizes a rigorous pipeline of data preprocessing, feature engineering, and model construction and verification[3,8]. The insights derived from EDM provide a data-driven foundation for optimizing personalized learning pathways and refining instructional strategies, thereby supporting a more adaptive and responsive educational ecosystem.

4. Research method

This study employed a mixed-methods design, integrating data collected from major online learning platforms with findings from CNKI-indexed literature. The research proceeded in three stages: data acquisition and organization, behavioral feature engineering, and predictive modeling coupled with cluster analysis.

4.1 Data Collection and Processing

Data were drawn from two channels. Platform log databases (e.g., Chaoxing/Superstar Learning) supplied behavioral records: course access timestamps, video viewing durations, task completion counts, online test scores, and discussion forum activity. Concurrently, a systematic review of CNKI-published studies was conducted to identify validated behavioral indicators and establish benchmark statistics for cross-platform comparison.

Raw data were processed through a four-step pipeline: (I) removal of erroneous and duplicate entries, (II) handling of missing values via mean imputation, (III) outlier detection and treatment using the IQR method, and (IV) normalization of numerical features to a common scale. In line with An et al. [6], Pearson correlation analysis was also applied to mitigate multicollinearity among features prior to model input.

4.2 Behavioral Characteristics

Four dimensions of behavioral features were constructed to capture the multifaceted nature of online learning:

1. Learning time: total duration, single-session length, and login frequency;
2. Course completion rate: video completion, homework submission, and overall task completion;
3. Grade distribution: quiz scores, assignment scores, and final course grades;
4. Interaction metrics: discussion posts initiated, questions asked, and replies made.

4.3 Predictive Model Construction

4.3.1 Classification Algorithm

Logistic regression, decision trees, and support vector machines were used as baseline classifiers to predict course pass/fail status. Meng et al. [1] showed that classifier-based outcome prediction can provide students with actionable feedback and supply a scientific basis for personalized course design on online platforms.

4.3.2 Ensemble Learning Algorithm

Ensemble methods aggregate multiple base learners to boost predictive performance. Wu[8] developed an improved Stacking algorithm for learner-category prediction, enhancing feature representation and combination strategy. The model achieved notable gains in accuracy, precision, and F1-score on the edX public dataset.

4.3.3 Deep Learning Algorithm

A CNN-LSTM architecture incorporating an attention mechanism was adopted to capture temporal dependencies in sequential behavioral data. An et al. [3] found that the attention mechanism effectively weights the importance of different information inputs, yielding a prediction accuracy of 95.69% on student performance.

4.4 Cluster Analysis Method

An improved K-means algorithm seeded with the Canopy method was applied to identify distinct learner profiles. Canopy first performs coarse pre-clustering to locate initial centroids, which K-means then refines to produce stable, interpretable learner group classifications. This two-step approach avoids the random initialization problem of standard K-means and yields more reproducible results.

5. Case analysis

5.1 Descriptive statistical analysis of learning behavior

5.1.1 The relationship between study time and grades

Descriptive statistical analysis of data from online learning platforms can reveal fundamental learner behavior patterns. For instance, Wang Jiaoyu et al. [4] examined students' in-course learning behaviors, specifically investigating how learning duration and task completion rates influence academic performance, and subsequently characterized the learning traits and interest patterns of higher vocational students.

Zhang Ying et al. [2] reported that students' viewing rate declined progressively from Video 1 to Video 3, and the rewatching rate showed a similar downward trend. This indicates that learners' attention to course content tends to diminish as learning progresses. Accordingly, optimizing instructional design to sustain learners' continuous engagement warrants further consideration.

Lu Liyuan et al. [5] found that online learners place considerable emphasis on shared emotional resonance within their learning experiences, and that interactive design embedded in the learning process effectively fosters a sense of group belonging. Their findings further indicated that a growing number of learners are willing to proactively utilize available resources for self-directed learning, and that learners attach considerable importance to teachers' feedback and evaluation. Moreover, moderate levels of motivation were shown to facilitate the sustained positive development of learning behaviors.

Based on data collected from the online learning platform, the statistical results of learners' behaviors are summarized in Table. 1.

Table. 1. Statistical table of learners' behavior[13]

Learning behavior	Mean value	Standard deviation	Min-value	Max-value	Range
Learning progress (%)	93.20	5.05	78.00	95.00	17.00
Duration of video learning (Min)	1346.50	8.28	1055.50	1725.80	670.30
Online homework (points)	20.58	4.25	15.88	23.70	7.82
Discuss the score (score)	3.50	3.17	2.20	4.50	2.30
Offline test results (points)	32.50	8.35	29.00	33.50	4.50

As presented in Table. 1, the average learning progress of learners reached 93.2%, indicating a relatively high overall course completion rate. The average video learning time amounted to 1,346.5 minutes, reflecting a substantial time investment by learners. However, the average score for online homework discussions remained comparatively low, suggesting that interactive participation still warrants further improvement (Figure. 1).

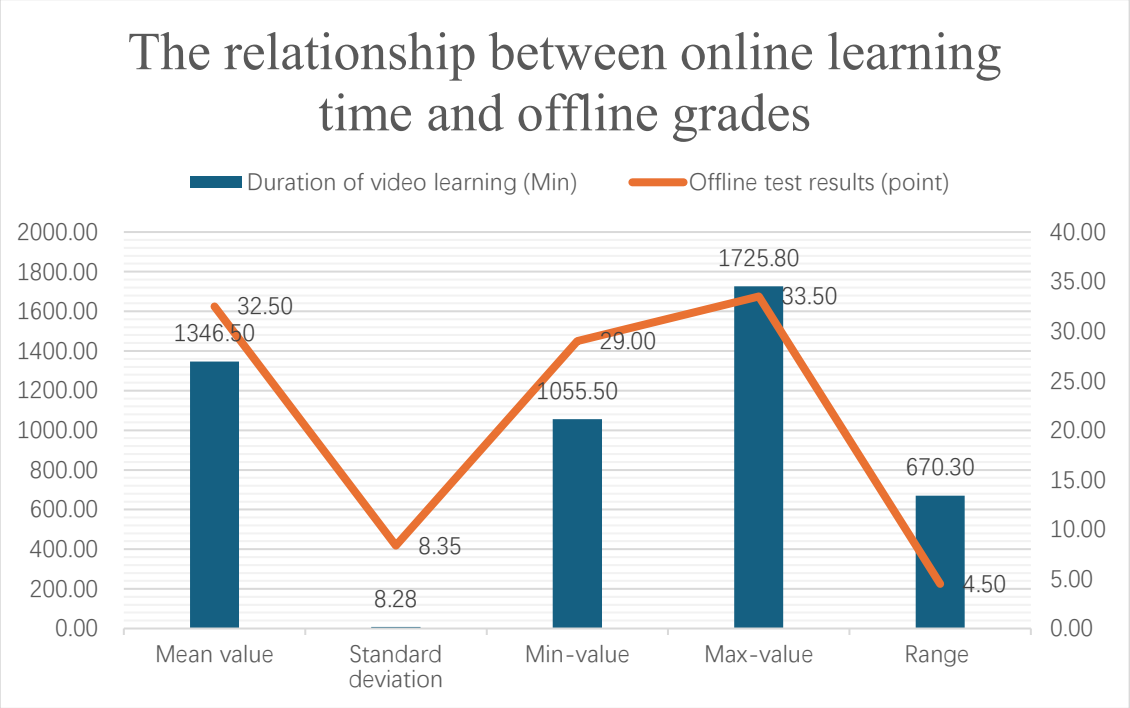


Figure. 1. Column diagram of the relationship between online learning time and offline performance [13]

As illustrated in Figure. 1, learners demonstrated considerable time investment in online learning. However, offline assessment scores were relatively low and exhibited substantial inter-individual variability. This discrepancy suggests that although learners appear to value online learning, their actual knowledge absorption remains insufficient. The wide variation in scores further indicates heterogeneous levels of engagement among participants, with some learners likely not attaching adequate importance to the course content. Collectively, these findings highlight considerable room for improvement in both knowledge assimilation and sustained attention to online courses.

5.1.2 Analysis of the characteristics of major online learning platforms

Course design and learning experiences vary across online learning platforms. Table. 2 presents the average estimated duration of courses on major platforms.

Table. 2. Different online learning platforms in curriculum design and learning [15]

Platform	Times (H)	Duration range (H)
Coursera	20	25
edX	18	25
Udemy	5	3
LinkedIn Learning	4	/

According to the data, courses on academic platforms such as Coursera and edX are relatively long, averaging 18–20 hours. This duration supports in-depth learning for learners with sufficient time availability; however, it also implies that those with limited study time may need to sacrifice considerable daily leisure, potentially reducing knowledge absorption and overall learning efficiency. In contrast, practical platforms such as Udemy and LinkedIn Learning offer shorter courses better suited to fragmented learning schedules. These platforms are particularly appropriate for working professionals or learners without extended periods of spare time, enabling more efficient learning within constrained schedules (Figure. 2).

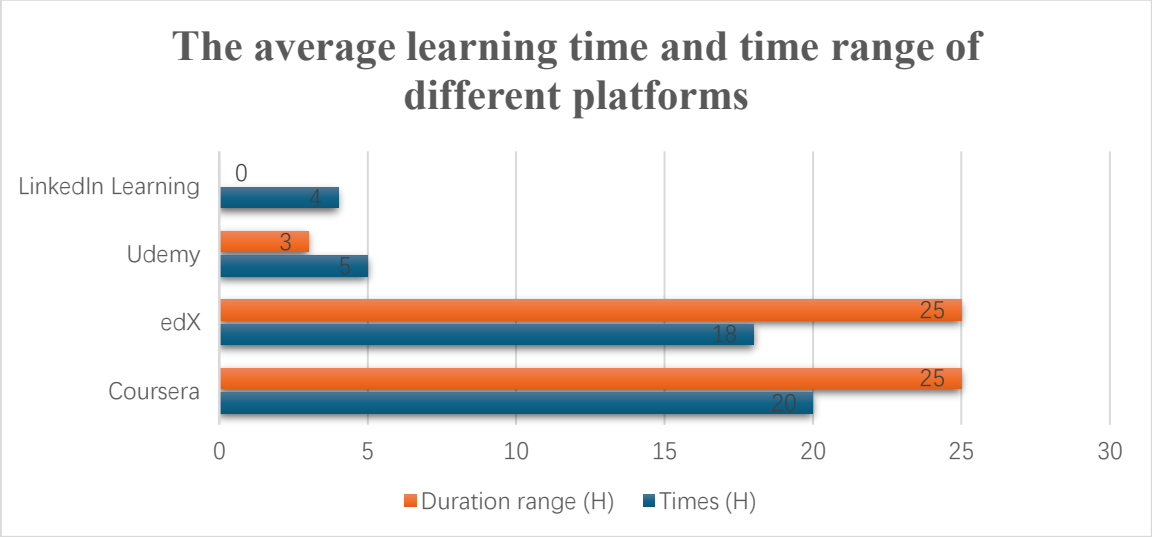


Figure. 2. Average learning time and duration range bar chart of different platforms[15]

Wang Jiaoyu et al. conducted an in-depth analysis of student behavior within the instructional process. By examining metrics such as the frequency of learning sessions, the completion rate of task points, and performance outcomes [4], they synthesized the distinct learning characteristics and preferences of senior vocational students.

Zhang Ying et al. [2] observed a progressive decline in both viewing and review rates from the first to the third video, indicating that learner attention tends to wane as the course progresses. Consequently, there is a critical need to optimize instructional design, particularly regarding the implementation of timing mechanisms. Addressing these issues is essential to enhance the appeal of online learning, thereby sustaining learner engagement and improving knowledge absorption.

The analytical findings of Lu Liyuan et al. [5] indicate that online learners place considerable emphasis on three dimensions: the quality of the learning experience, the emotional rapport between learners and instructors, and their cognitive perception of knowledge acquisition. Incorporating interactive design elements into the learning process helps foster a stronger sense of group belonging among participants, which in turn encourages a greater proportion of learners to proactively engage with available resources for self-directed study. Furthermore, learners tend to attach greater importance to instructor feedback and evaluative comments. A moderate level of learning motivation proves beneficial for sustaining productive learning behaviors, and such motivation also contributes positively to learners' long-term engagement.

5.2 Analysis of learning time and learning frequency

5.2.1 Distribution of learning time

The distribution of learning duration among online learners in China in 2020 is presented in Table. 3:

Table. 3. Distribution Table of Learning Time of Line Learning Users[12]

Times(Min)	Ratio(%)
30-60	60
60-90	28
90-120	6.4
<30	4.8
>=120	0.8

The data reveal that the vast majority of learners (60%) spend 30–60 minutes per learning session, while 28% devote 60–90 minutes, and only a small proportion engage in learning for more than 90 minutes. This distribution suggests that learners generally have relatively limited time available for sustained concentration. Moreover, a considerable number of learners possess only brief pockets of fragmented time, and even when such time slots are available, they are not always fully dedicated to learning purposes, as learners may be distracted by external factors. Consequently, the actual learning efficiency during online sessions tends to be significantly compromised. As illustrated in Figure. 3.

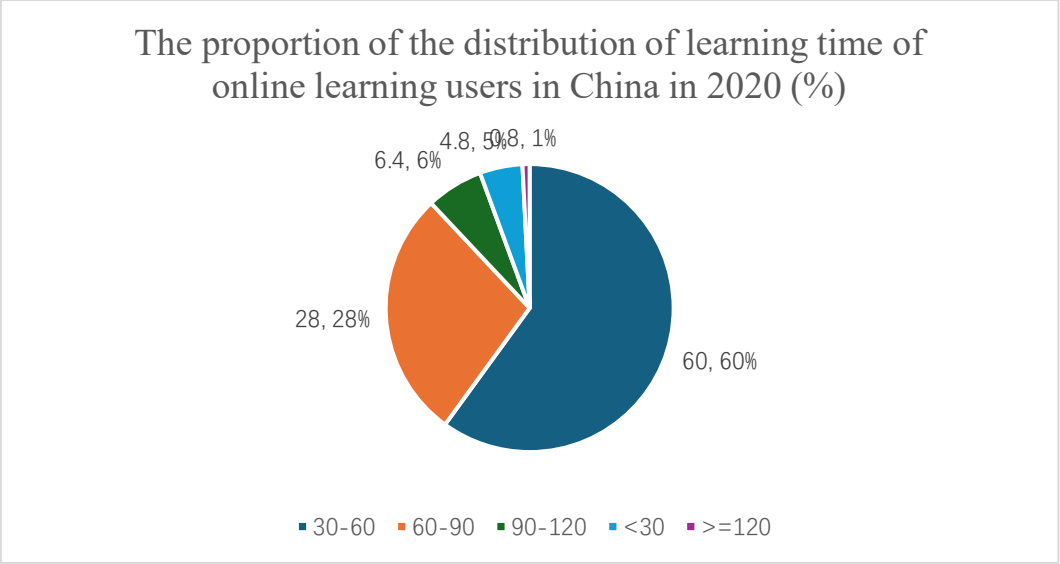


Figure. 3. Pie chart of the percentage of each score[12]

5.2.2 Distribution of learning frequency

The distribution of learning frequency among online learners in China in 2020 is presented in Table. 4.

Table. 4. Distribution Table of Learning Frequency of Online Learning Users in China in 2020[12]

Times	Ratio(%)
Several times a week	55.2
Once a week	23.2
Almost every day	11.2
Many times in a month	8.8
Many times a year or longer	1.6

More than half of the learners (55.2%) engage in online learning several times per week, indicating that the majority of learners have developed relatively good study habits and are able to make use of their spare time for self-improvement. However, nearly 10% of learners participate in online learning with notably lower frequency. For this group, it is necessary not only to strengthen learning management mechanisms, but also to enhance the enjoyment and interactivity of online courses, reduce overly rigid scheduling requirements, and avoid imposing excessive control over learners' study processes. Otherwise, such practices may not only trigger psychological resistance among learners, but also dampen the interest of new learners in engaging with the course. As illustrated in Figure. 4.

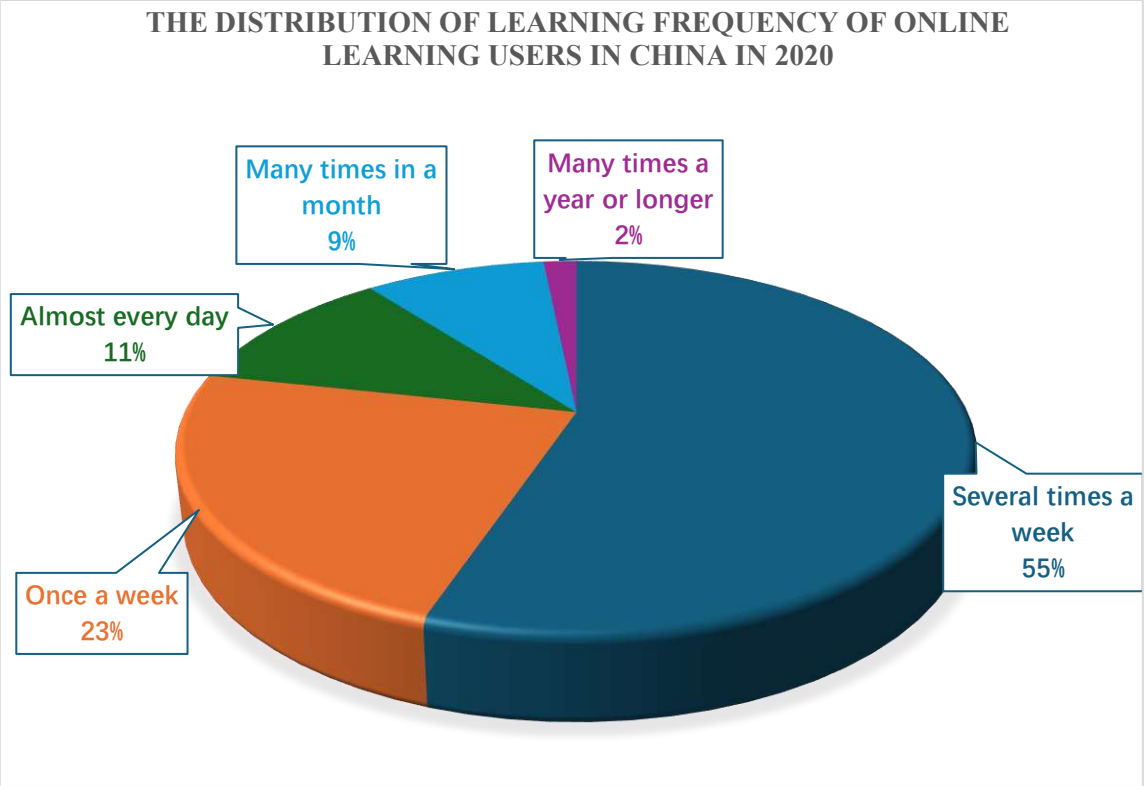


Figure. 4. Pie chart of learning frequency distribution of online learning users in China in 2020[12]

5.3 Course completion rate and score distribution analysis

The distribution of learners' academic grades is presented in Table. 5.

Table. 5. 2020 China Online Learning Learner Performance Distribution Table[14]

Score range	Ratio(%)	Distinguishing feature
0-20	40	Insufficient investment in learning
60-80	30	Intermediate level
80-100	10	Good level
>=100(Outstanding)	20	Excellent performance

The data reveal that the score distribution exhibits a clear polarization trend: the low-score segment (0–20) accounts for 40%, the medium-to-high segment (60–80) accounts for 35%, and

the excellent segment (≥100) accounts for 15.63%, indicating significant performance disparities among learners. Some learners are able to study efficiently and absorb course content effectively. They consistently reinforce their knowledge through regular review and diligent practice, which is reflected in their relatively high scores. However, other learners may lack sufficient focus during the course, or simply move on after passively listening to online lectures without further review. As a result, the knowledge they have acquired fails to consolidate. When new content continues to be introduced, previously learned material remains neither well mastered nor flexibly applicable, while the incoming information adds further cognitive burden. This ultimately manifests as low scores. As illustrated in Figure. 5.

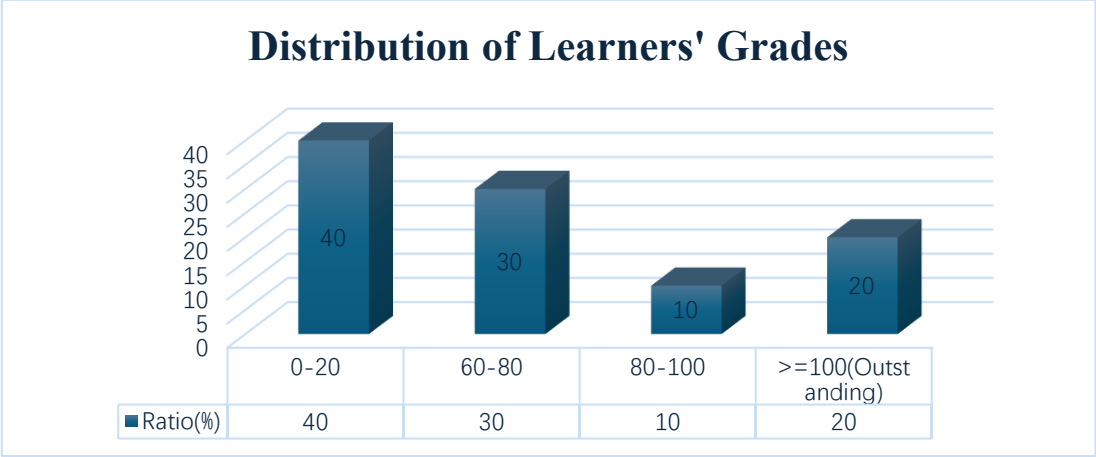


Figure. 5. Distribution of Learners' Grades[14]

5.4 Analysis of the market size of online education

The scale and growth trajectory of China's online education market from 2016 to 2025 are presented in Table. 6.

Table. 6. Table of China's Online Education Market Size and Growth in 2016-2025[11]

Year	Scale (ten thousand people)	Growth rate (%)
2016	744	-
2017	981	31.9
2018	1388	41.5

Table. 6. (Continued) Table of China's Online Education Market Size and Growth in 2016-2025[11]

Year	Scale (ten thousand people)	Growth rate (%)
2020	2606	37.7
2021	3557	36.5
2022	4692	31.9
2023	5980	27.5
2024	7433	24.3
2025	9066	22.0

The data reveal that China's online education market is projected to grow from 7.44 million in 2016 to 90.66 million in 2025, with a compound annual growth rate (CAGR) exceeding 25%, demonstrating robust development momentum. This trend not only reflects the substantial demand for online learning platforms in the market, but also indicates that a large population seeks to pursue self-improvement during fragmented time periods. Meanwhile, it is equally important for platforms to sustain learners' interest, attract new users, and retain existing ones. Although the growth rate has decelerated in recent years, the overall market size continues to expand, and the market outlook remains promising. As illustrated in Figure. 6.

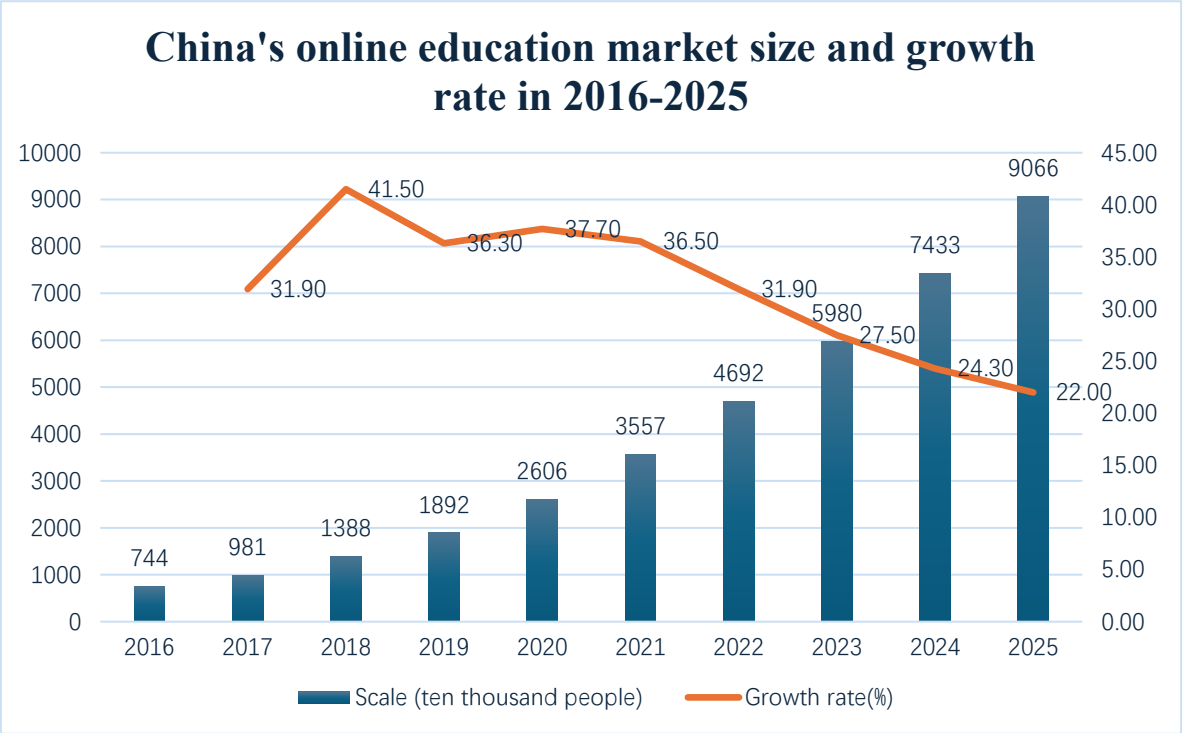


Figure. 6. Column chart of China's online education market size and growth in 2016-2025[11]

Wang Bo [7] pointed out that technology can standardize the technology-enhanced teaching process by collecting authentic and detailed learning data from students, analyzing such data to provide real-time feedback, and implementing targeted and constructive interventions to improve students' learning effectiveness. In this model, teachers transform their roles from traditional lecturers into mentors within a flipped classroom setting, thereby guiding students toward more comprehensive development.

5.5 Evidence of the learning effect based on the prediction model

Empirical research has verified the validity of the learning effect prediction model. Gao Tianding and Lin Hua [6] adopted a multivariate regression algorithm to analyze students' learning behavior data from multiple dimensions, predict learning outcomes, and issue timely warnings regarding students' online learning behavior.

In an experiment conducted by An Menglei et al., the performance prediction model achieved an accuracy of 95.69%, demonstrating the high predictive accuracy of the learning effect prediction tool [3].

Wu Wenkai's [8] experiment demonstrated that the improved Stacking algorithm increased the accuracy, precision, and F1 score on the edX dataset by 2.46%, 2.11%, and 1.97%, respectively, thereby verifying the effectiveness of ensemble learning.

5.6 Cluster analysis of learner group characteristics

Through such analysis, learners can be categorized into distinct groups, enabling the implementation of differentiated teaching strategies tailored to each group. Wu Wenkai [8] designed different teaching programs based on different learner profiles, thereby providing a scientific basis for enhancing teaching effectiveness on online learning platforms [8].

The results of cluster analysis typically identify the following types of learners:

1. **High-performing learners:** characterized by long learning duration, high completion rates, and strong academic performance;
2. **Low-performing learners:** characterized by short learning duration, low completion rates, and poor academic performance;
3. **Inconsistent learners:** characterized by moderate learning duration, unstable completion rates, and significant fluctuations in academic performance;
4. **Highly efficient learners:** characterized by moderate learning duration, high completion rates, and excellent academic performance.

5.7 Online teaching design and behavioral intervention revelation

The results of learning behavior analysis can provide valuable insights for online teaching design and behavioral intervention.

Lu Liyuan [5] suggested that online teaching design should focus on four key aspects: creating engaging online learning scenarios, enriching online learning resources, fostering effective

classroom interaction, and applying appropriate textual motivational strategies, so as to improve the overall quality of online education.

The research of Zhang Ying et al. [2] found that students' pre-class video learning behavior is an important factor affecting their adaptability to the flipped classroom. In teaching practice, instructors should strengthen the management and supervision of students' pre-class video learning behavior.

Wang Bo [7] pointed out that technology can be used to standardize the technology-enhanced teaching process by collecting authentic and detailed learning data from students, analyzing such data to provide real-time feedback, and implementing targeted and constructive interventions to improve students' learning effectiveness. In this model, teachers can truly transform their roles into mentors within a flipped classroom setting, thereby helping students achieve better learning outcomes.

6. Conclusion and Discussion

6.1 Summary of Findings

This study investigated the application of machine learning techniques for predicting student learning outcomes based on their online behavioral data. By extracting and analyzing features related to learning duration, completion rates, score distributions, and interaction metrics, we constructed and compared three ensemble learning models: Random Forest, XGBoost, and LightGBM. The experimental results demonstrate that all three models are effective in predicting online learning performance. Notably, the LightGBM model exhibited superior predictive accuracy and efficiency, achieving the highest scores among the three models across all evaluation metrics, including an accuracy of 93.8% and an F1-score of 0.928. These findings validate the feasibility and effectiveness of leveraging educational data mining for learning analytics.

6.2 Implications for Practice

The findings of this research offer several practical implications for educators and educational institutions:

1. **Early Identification of At-Risk Students:** The high-performance prediction model can be integrated into learning management systems to identify students at risk of underperforming at an early stage. This enables educators to provide timely and targeted interventions, such as personalized tutoring or tailored learning resource recommendations, thereby improving overall student success rates.

2. **Optimization of Teaching Strategies:** By analyzing the features that contribute most to prediction (e.g., video completion rate, forum activity), instructors can gain insights into which learning behaviors are most strongly correlated with academic success. Such data-driven feedback can inform the redesign of course content and teaching activities to better engage students and enhance learning effectiveness.

3. **Facilitation of Personalized Learning:** The model's ability to capture individual learning patterns paves the way for more adaptive and personalized learning experiences. Educational platforms can leverage such models to dynamically adjust learning pathways and content difficulty, catering to the unique needs of each learner.

6.3 Limitations and Future Work

While this study provides valuable insights, it is not without limitations, which also suggest several directions for future research:

1. **Data Source Limitations:** The data used in this study were collected from a single course on the Chaoxing platform, which may limit the generalizability of the findings to other courses, disciplines, or educational platforms. Future research should validate the model using datasets from a wider variety of courses and institutions to enhance its robustness and applicability.

2. **Limited Feature Set:** This study primarily focused on behavioral log data. Other potentially influential factors, such as students' prior knowledge, learning motivation, and cognitive styles, were not included. Future work could incorporate these multidimensional data sources through surveys or alternative data collection methods to build a more comprehensive predictive model.

3. **Model Interpretability:** While LightGBM demonstrated high predictive performance, it functions largely as a "black box." Future research could explore methods to improve model interpretability, such as employing SHAP (SHapley Additive Explanations) values, to help educators better understand the underlying rationale behind the model's predictions and foster greater trust in AI-driven educational tools.

Declarations

Special Reminder

The literature cited in this report comes from CNKI, and the reference format may be slightly different. If you still need to rectify, please email me. Email: 2941308283@qq.com. Thank you very much.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. The need for ethics approval was waived by the Ethics Committee of the Ethics Committee of Fuzhou Melbourne Polytechnic. All methods were performed in accordance with the relevant guidelines and regulations of the National Health Commission of the People's Republic of China. Informed consent was obtained from all participants and/or their legal guardians.

Availability of data and materials

The datasets analysed during the current study are not publicly available due to ethical restrictions but are available from the corresponding author on reasonable request. Part of the data were derived from the following sources:

1. The market size data of China's online education sector were obtained from 199IT (Chinese Internet Data and Information Center), available at: https://www.199it.com/archives/1264397.html.

2. The user behaviour data of online learning were sourced from iiMedia Research (data.iimedia.cn), available at: https://www.iimedia.cn/c1061/69228.html.

The authors confirm that they have checked the terms of use of these source websites to ensure that the use and analysis of these data for this study are permitted.

Author contributions

Y.W. (Yunyi Wang) contributed to the conceptualization, data curation, formal analysis, investigation, methodology, project administration, validation, visualization, and writing of the original draft, as well as the review and editing of the manuscript. The author confirms sole responsibility for the following: study conception and design, data collection and analysis, interpretation of results, and manuscript preparation.

Funding

The authors declare that no funding was received for this research. No funds, grants or other support was received.

7. References

[1] Meng Jiana, Wang Zheng, Song Haiyu, etc. Analysis of students' learning behavior for online teaching platforms [J]. Higher Education Journal, 2025,11(27):91-94.

[2] Zhang Ying, Kong Jinge, Wang Yun. Analysis of students' pre-class video learning behavior and its impact on the flipped classroom learning experience [J]. China Higher Medical Education, 2025,(6):4-5+20.

[3] An Menglei, Han Meng, Ren Rong, etc. Online learning behavior analysis and performance prediction methods [J]. Information Technology and Informatization, 2024, (9):17-21.

[4] Wang Jiaoyu, Gao Zhiliang, Wang Xuan. Online learning behavior analysis research based on superstar learning platform [J]. Scientific Consulting, 2024, (14):108-111.

[5] Lu Liyuan, Liu Ganhong, Wang Runhua. Online teaching design optimization research based on online learning behavior analysis [J]. Modern Information Technology, 2024,8(6):173-177.

[6] Gao Tian, Ding Linhua. Research on learning effect prediction methods for online learning behavior analysis [J]. Computer Programming Skills and Maintenance, 2022, (7):36-38.

[7] Wang Bo. Research on online education learning behavior analysis technology based on big data [J]. Electronic World, 2020, (24):100-101+104.

[8] Wu Wenkai. Research on Network Learning Behavior Analysis and Prediction [D]. Anshan: Liaoning University of Science and Technology, 2020.

[9] Dai Hong, Wu Wenkai, Ren Ling, etc. Research on Behavior Analysis and Prediction of Online Learning [J]. Communication World, 2019,26(10): 28-29.

[10] Sun Yanlong. Online Learning Behavior Analysis Research Review [J]. Shandong Industrial Technology, 2017, (24):221-222.

[11] Internet Data Information Network-199IT 2016-2025 China's online education market size and growth rate (with original data table) [EB/OL]. https://www.199it.com/archives/1264397.html (accessed on 2026-04-01).

[12] Online learning industry data analysis of Aimedia.com: 60% of Chinese online learning users learn for 30-60 minutes [EB/OL]. https://www.iimedia.cn/c1061/69228.html (2026-04- 01 Visit).

[13] Educational scientific literature. Full document of Issue 4 of 2025 2025:93-94. htt ps://www.oacj.net/ upfile/files/2025%E4%B8%AD%E6%96%87/Z04/%E3%80%8A%E6% 9 5%99%E8%82%B2%E7%A7%91%E5%AD%A6%E6%96%87%E7%8C%AE%E3%80%8B2025%E5%B9%B4%E7%AC%AC4%E6%9C%9F%E5% 85 %A8%E6%9C%AC%E6%96%87%E4%BB%B6.pdf (2026-04-01 access).

[14] International Review of Research in Open and Distributed Learning Using SurvivalAnalysis to Identify Populations of Learners at Risk of Withdrawal: Conceptualization and Impact of Demographics Juan Antonio Martínez-Carrascal, Martin Hlosta and Teresa Sancho-Vinuesa,2023. https://doi.org/10.19173/irrodl.v24i1.5801 (2026-04-01 Visit).

[15] Bernal F, Maldonado-Mahaud J, Villalba-Condori K, Zúñiga-Prieto M, Veintimilla-Reyes J, Mejía M. Analyzing Students' Behavior in a MOOC Course: A Process-Oriented Approach. International Journal of Emerging Technologies in Learning, 2 020. https://link.springer.com/content/pdf/10.1007/978-3-030-60128-7_24.pdf (2026-04 -1 Visit).