Dean&Francis
ISSN 2959-6130

Analysis of Factors Affecting Academic Performance - Based on Machine Learning Methods

Zilin Liu

School of Economics and Management, Shihezi University, Shihezi, Xinjiang, China
20231016094@stu.shzu.edu.cn

Abstract:

Students' academic performance is the core indicator of educational quality. To overcome the limitations of traditional methods in analyzing the interaction and dynamic effects of complex factors, this study employs machine learning (logistic regression, ensemble learning, neural networks, SHAP value analysis, etc.) to integrate multi-source educational data and systematically explore the influencing factors and mechanisms of academic performance. Key findings: (1) Individual initiative (motivation, self-efficacy, strategy) contributed the most (48%), which was higher than that of teaching management (32%) and environmental factors (20%); (2) Motivation and self-efficacy have a synergistic enhancing effect, while working more than 30 hours per week weakens the role of the knowledge base. (3) There are significant differences among disciplines (mathematics emphasizes cognitive strategies, while medicine is regulated by psychological states). The research breaks through the static attribution paradigm, providing data support for precise teaching and pointing out the future direction of deepening dynamic modeling and interdisciplinary integration.

Keywords: Academic performance, influencing factors, machine learning, learning motivation

1 Introduction

Student academic performance, as a core indicator for measuring educational quality, requires the precise identification and analysis of its influencing factors have significant practical significance for optimizing teaching strategies, promoting educational equity, and enhancing the effectiveness of talent cultivation. In the era of the knowledge economy, the quality of higher education is directly related to a country's innovation capacity and core competitiveness [1]. A deep understanding of the driving and hindering mechanisms behind achievements is a key prerequisite for achieving personalized education and effectively implementing "teaching students by their aptitudes".

The current research background indicates that the factors influencing students' academic performance present complex characteristics such as multi-di-

ISSN 2959-6130

mensionality, interactivity, and disciplinary differences. A large number of empirical studies have been conducted to explore from multiple aspects, such as individuals, teaching, and the environment. Personal factors have been generally confirmed to occupy a core position, such as the learning foundation (for example, the influence coefficient of the knowledge reserve upon enrollment on the course grades of the pharmacy major is as high as 0.65 [2]. Admission scores significantly affect the mathematics scores of English major students [3]. The correlation coefficient between the previous academic performance and the subsequent learning effect reached 0.6-0.7, and the correlation coefficients between learning motivation and self-efficacy (the challenge and enthusiasm dimensions of endogenous motivation and academic performance reached 0.68 and 0.72, respectively [4, 5]. The interweaving of internal and external motivations affects the English scores of vocational undergraduate students [1]. Learning interest and self-efficacy have a significant impact on junior high school mathematics scores as well as learning strategies and time investment (the influence coefficient of learning methods is 0.72 [6, 7]. The influence coefficient of the weekly self-study duration is 0.58 [7]. Learning habits have a significant positive impact on the grades of advanced mathematics in the first year of college [8]. Psychological states (anxiety and depression are significantly negatively correlated with the grades of some specialized courses) also play important roles [7]. Secondly, the role of teaching and management factors is also very significant, covering teaching methods and resources (the influence coefficient of heuristic teaching on the first-year advanced mathematics score is 0.62 [8]. The influence coefficient of the richness of teaching resources is 0.45 [7]. The influence coefficient of teachers' teaching methods on junior high school mathematics scores reached 6.332, the curriculum setting and its correlation (such as suggestions for the construction of related course clusters), as well as the academic atmosphere and facility guarantee (with influence coefficients of 0.55 and 0.48, respectively) [6, 7]. In addition, environmental and background factors cannot be ignored. This includes work burden (students who work more than 30 hours per week have a significantly lower average academic performance), family environment (students with harmonious parents and a shared focus on learning have higher academic performance), grade differences and professional perception (the attention to employment prospects has an opposite impact on math grades across different grades), and gender differences (in specific subjects and courses There is a significant impact [6-8].

Although the above-mentioned research employed traditional statistical methods such as correlation analysis, regression models (linear, Logistic), and analysis of variance and achieved rich results, laying a foundation for understanding the influencing factors of performance, there are still significant gaps in current research. Firstly, there are limitations in methods: traditional statistical methods are relatively insufficient in handling high-dimensional features, nonlinear relationships, and complex interaction effects, making it difficult to fully explore the hidden patterns and rules in the data. Secondly, there is a lack of dynamism and predictability: Most studies focus on static correlation analysis and lack dynamic prediction models based on historical data, which cannot provide effective tools for early academic warning and precise intervention. Finally, the detailed characterization of factor weights and interactions is insufficient: The existing research still needs to deepen the precise quantification of how numerous factors work together and their relative importance, making it difficult to provide clear and actionable priority guidance.

Therefore, this research holds significant theoretical value and practical significance: At the theoretical level, the introduction of machine learning methods aims to break through traditional limitations and deeply reveal the influence mechanism and interaction network of academic performance; At the practical level, building high-precision predictive models can provide data support for educational decision-making and facilitate personalized intervention and resource allocation.

This study intends to adopt machine learning (ensemble learning, neural networks, SHAP value analysis, etc.) to integrate multi-dimensional data and systematically analyze the key influencing factors of performance and their action paths. The specific contents include: (1) Constructing and evaluating the performance prediction model; (2) Contribution of quantitative factors; (3) Exploring the interaction effects among factors; (4) Comparing the differences in subjects and grades.

The core objective is: (1) To verify the superiority of machine learning in parsing performance factors; (2) Identify the core variables and their mechanisms of action; (3) Provide a scientific basis and tools for precise teaching. The background and current situation, research methods, analysis results, and discussion suggestions will be elaborated in sequence subsequently.

2 The Evolution of performance prediction methods in the field of Education

2.1 Limitations of Traditional Mathematical Statistics Analysis Methods

The current empirical research on education mainly relies on qualitative analysis and mathematical statistics models, the latter of which generally employs techniques such as

correlation analysis, principal component analysis, and regression analysis. Although these methods have a clear theoretical framework in the exploration of variable relationships, there are significant constraints on model applicability. Classical linear regression based on a continuous variable requires that the dependent variable satisfy a normal distribution and there be no multicollinearity among the independent variables [6]. When the dependent variable is categorical data, complex variable transformation or information loss (such as forcibly quantifying grade levels) is needed. Although the problem of categorical dependent variables (such as excellent/good/medium/poor in advanced mathematics) was solved through ordered multiclass Logistic regression, its model is still limited by the proportional advantage assumption (parallelism test) and cannot effectively capture nonlinear interaction effects.

Specifically, bivariate regression, multiple linear regression, and logistic regression constitute the core methodological system of educational empirical research. Bivariate regression quantifies the causal trends of two variables through a univariate linear equation (y=β0+β1, xy=β0+β1, x). Its advantage lies in its simplicity and intuitiveness. For instance, confirms that for every 1-point increase in English scores, math scores increase by 0.865 points (β1=0.865, t=7.947; β1=0.865, t=7.947) [9]. Multiple linear regression is extended to a multi-independent variable model (y=β0+∑βixi, y=β0+∑βixi), which has a remarkable ability to control confounding variables and analyze independent contributions [6]. Based on this, it is determined that the influence of learning interest (β=12.298, β=12.298) on mathematics grades far exceeds that of teaching methods (β=6.332, β=6.332). Logistic regression processes the categorical dependent variable through Logit transformation (ln(p1-p)=β0+∑βixiln(1-pp)=β0+∑βixi), and the characteristic of output probability prediction applies to the grading system score analysis [7]. It was found that the independent completion of homework significantly increased the probability of excellent grades in advanced mathematics (β=3.685, β=3.685). Three types of methods form a progressive analytical framework: bivariate regression preliminarily verifies the association, multiple regression reveals the role of multiple factors, and logistic regression solves the prediction of discrete results, jointly supporting the scientific and refined nature of the research on educational influencing factors.

The deeper limitations are reflected in the data processing dimension: traditional methods rely on manually preset variable structures, making it difficult to handle high-dimensional unstructured data (such as learning behavior texts, classroom videos, etc.) [3]. In terms of variable interaction analysis, none of the above-mentioned studies examined the complex correlations among factors, while principal component analysis can reduce dimensions, it comes at the cost of losing the explainability of the original features. Furthermore, these methods are sensitive to sample size and distribution. For instance, the extrapolation of the t-test results of a 67-person sample is questionable, and its linear regression model (English → mathematics scores) does not control for confounding variables (such as cognitive ability), which can easily lead to pseudo-correlation [3].

2.2 Application Potential of Machine Learning Methods

Machine learning methods, through adaptive feature engineering and nonlinear modeling capabilities, provide a new paradigm for educational prediction research. Compared with the strict restrictions on variable types imposed by traditional statistical models, machine learning algorithms can flexibly handle mixed-type data: Decision Tree and its integrated methods (such as Random Forest) can directly parse categorical variables (such as job plagiarism =1/ independent completion =4), and identify key influencing factors through feature importance ranking. Support Vector Machine (SVM) can fit complex decision boundaries through kernel function transformation, effectively solving the possible implicit nonlinear clustering problem of insignificant class difference [3].

At the level of model performance, machine learning demonstrates three advantages: Firstly, Neural Networks automatically extract high-order interaction features (such as the synergistic impact of "dormitory academic atmosphere × teacher Q&A frequency" on grades) through a multi-layer perception mechanism, overcoming the defect that Logistic regression requires manual setting of interaction terms; Secondly, integrated methods (such as XGBoost) have stronger robustness against high-dimensional sparse data (such as tens of millions of behavior logs in student digital portraits), and can avoid the problem of information loss in principal component analysis. Finally, clustering algorithms (such as K-means) can unsupervised identify potential student groups (such as the "high motivation - low foundation" group), providing a detailed perspective on the heterogeneous influence of "non-intellectual factors" [6]. It is worth noting that deep learning models (such as LSTM) perform outstandingly in time series data analysis, capable of tracking the dynamic evolution of learning behaviors (such as the prediction of final grades based on the trend of assignment quality), which is something that cross-sectional regression models cannot achieve.

However, machine learning also faces interpretability challenges. The "black box" feature of random forests makes it difficult to directly output the OR value (Odds Ratio) of Logistic regression, and it relies on post-interpretation techniques such as SHAP values (SHapley Additive exPlanations). Future research needs to combine

ISSN 2959-6130

probabilistic graphical models such as Bayesian networks to enhance the causal inference ability of educational decisions while maintaining prediction accuracy.

3 Case Analysis

3.1 Analysis of Influencing Factors in Advanced Mathematics

In order to study the specific factors influencing advanced mathematics scores, Pan Xingxia et al. conducted an empirical study on the influencing factors of advanced mathematics scores and innovatively constructed a three-dimensional analysis framework for education, learning, and management [7]. The research adopted the structured questionnaire survey method to collect data from 800 freshmen in 11 colleges of Nanchang Hongkong University. After eliminating invalid samples through contradiction item verification, 668 valid questionnaires were retained (with an effective rate of 84%). In terms of variable design, the 26 influencing factors are classified into three major categories: personal factors (learning foundation/method/motivation/interest), management factors (academic atmosphere/facilities/interpersonal relationships), and teaching factors (content/interaction/teacher quality). The classification variables are standardized and assigned values based on the Likert scale (for example, homework completion method: plagiarism =1 to independent completion =4). To ensure data reliability, the study tested the reliability through the Cronbach Alpha coefficient (overall reliability 0.852) and verified the validity using the KMO-Bartlett method (KMO=0.865, Bartlett test $p<0.001$).

In terms of the analytical method, the dependent variable "advanced mathematics grade" (excellent/good/medium/poor) is an ordered categorical variable, so an ordered multi-class Logistic regression model (cumulative Logit model) is adopted. The model passed the parallel line test ($\chi^2=171.978$, $p=0.153>0.05$) to satisfy the proportional advantage hypothesis, and the variance inflation factor VIF<10 excluded multicollinearity interference. The results of the likelihood ratio test ($\chi^2=342.416$, $p<0.001$) indicated that the model was statistically significant.

Empirical results show that personal factors have the most significant impact on academic performance. Among them, completing homework independently has a positive effect compared to the plagiarism method (with a β coefficient difference of 3.685), while pure self-study (β=-0.719) or pure listening to classes (β=-0.497) are both inferior to the blended learning mode. Among the management factors, the number of students who failed in dormitories (β=-1.322) and the dissatisfaction with classroom facilities (β=-0.768) showed a significant negative impact. Among the teaching factors, the inappropriate difficulty of the teaching materials (β=-1.045) and the management of classroom order (β=1.066) play a prominent role, but the interactive indicators such as the frequency of teachers and students answering questions have not passed the significance test. It is worth noting that although students' self-reports emphasize the importance of personal factors, attribution analysis reveals that they are more inclined to attribute academic performance issues to external factors such as course difficulty (accounting for 57.3%), and only 12% admit that they have not made sufficient efforts.

This study effectively solved the problem of analyzing classification dependent variables through an ordered Logistic model. However, due to the limitations of a single institution sample and the lack of investigation into variable interaction effects, the universality of the conclusion needs further verification.

3.2 Research on Non-Intelligence Factors in Junior High School Mathematics

In order to study the specific influences of non-intellectual factors on junior high school mathematics scores, Zhao Huani et al. systematically investigated the influencing factors of non-intellectual factors on junior high school mathematics scores through the multiple linear regression method [6]. This study took 221 students from grades seven to nine of a middle school in Longdon City as samples and collected data by combining questionnaire surveys with examination results. Eventually, 193 valid questionnaires were obtained. The research focuses on three key non-intellectual factors: interest in mathematics learning, self-efficacy, and teachers' teaching methods, and reveals their influencing mechanisms through rigorous statistical analysis methods. In the data processing stage, the research first inspected the quality of the mathematics test papers, confirming that they had good discrimination (total discrimination 0.538) and moderate difficulty (total difficulty 0.719), providing a reliable data basis for subsequent analysis.

The study adopted the Kolmogorov-Smirnov test to conduct normality analysis on the data, and flexibly selected the analysis method based on the test results. For the mathematics learning interest data that do not conform to the normal distribution, a non-parametric test method was adopted. The results showed that the mathematics scores of students in the high-interest group (56.8±24.7) were significantly better than those in the low-interest group (24.4±17.9). For the data of self-efficacy and teachers' teaching methods that conform to the normal distribution, the independent sample t-test was used for analysis. It was found that the performance of students in the high-efficiency group (64.2±20.5) was significantly higher than that in the low-efficiency group (35.6±20.0). Moreover,

the students' scores in the high-frequency praise group by teachers (57.3±23.8) were significantly better than those in the low-frequency praise group (39.4±22.4). These differences were all statistically significant ($p<0.001$), confirming the important influence of non-intellectual factors on mathematics scores.

Based on the above findings, a multiple linear regression model was constructed in the study, and the reliability of the results was ensured through strict model checking. The variance inflation factor (VIF) was all less than 1.5, eliminating the influence of multicollinearity. The model significance test ($F=127.345$, $p<0.001$) and the goodness-of-fit index (adjusted $R^2=0.718$) indicated that the model had good explanatory power. The final regression equation obtained shows that mathematics learning interest, self-efficacy and teachers' teaching methods all have a significant positive impact on mathematics grades. Among them, the influence of learning interest is the most prominent ($\beta=12.298$), followed by self-efficacy ($\beta=8.614$) and teachers' teaching methods ($\beta=6.332$). This quantitative result provides a clear guiding direction for educational practice.

This study, through rigorous empirical analysis, confirmed the significant role of non-intellectual factors in junior high school mathematics learning, providing a scientific basis for educators to optimize teaching strategies. However, the research also has certain limitations, such as a single sample source and relatively simple measurement dimensions for teachers' teaching methods. All these need to be improved and refined in future studies. Overall, this study demonstrates normativity in methodology and emphasizes accuracy in result presentation, providing valuable references for research in related fields.

3.3 Correlation with the Mathematical Performance of English Majors

To study the factors influencing the mathematics performance of English major students, Ji Hongwei conducted research on the influencing factors of English major students' mathematics performance through statistical analysis methods [3]. The study took 67 students from Class 2 and Class 3 of the English major of the 2008 grade in a certain university as samples, focusing on three core issues: the correlation between admission scores and mathematics scores, the differences in scores among different classes with the same teacher, and the correlation between English and mathematics scores. In terms of methodology, the study employed an independent sample t-test to analyze the impact of admission scores. The admission scores were divided into the high group (≥80 points) and the low group (<80 points) based on the 80-point threshold. After confirming the homogeneity of variance by Levene's homogeneity of variance test ($F=2.955$, $p=0.090>0.05$), the t-test results showed a significant difference in the mathematics scores between the two groups ($t=4.442$, $p<0.001$). It indicates that the admission score has a significant impact on the mathematics score (the average mathematics score of the high group is 7.794 points higher than that of the low group).

To address the issue of performance differences among different classes under the same teacher's instruction, one-way analysis of variance (ANOVA) was used for testing. The results showed that there was no significant difference between classes ($F=1.112$, $p=0.296>0.05$), supporting the null hypothesis that "class differences are not related to teachers", indicating that teaching consistency can effectively control the influence of class variables on academic performance.

For the relationship between English and mathematics scores, the research adopted a method combining bivariate correlation analysis and univariate linear regression. Pearson correlation analysis showed a moderate positive correlation between the two subjects' scores ($r=0.702$, $p<0.001$). Subsequently, a regression model was established with English scores as the independent variable and mathematics scores as the dependent variable. Analysis of variance confirmed the significance of the model ($F=63.156$, $p<0.001$), and the regression coefficient was 0.865 ($t=7.947$, $p<0.001$), resulting in the equation: Mathematics score $=14.402+0.865\times$ English score. The results show that for every 1-point increase in English scores, the expected increase in mathematics scores is 0.865 points, confirming the promoting effect of language ability on mathematics and science learning.

This study verified the empirical assumptions in educational practice through rigorous statistical methods. However, due to the sample size ($n=67$) and the lack of control for confounding variables such as learning motivation, the extrapolation of conclusions should be approached with caution. Its methodological value lies in demonstrating the collaborative application logic of t-tests, ANOVA, and regression analysis in educational empirical research [10].

4 Discussion

Table 1 summarizes key findings and limitations from three educational studies.

ISSN 2959-6130

Table 1. Summary of Selected Educational Research Cases and Their Findings

Case	Sample/Method	Significant Findings	Limitations
[7]	668 college students; Ordered logistic regression	(1) Independently completing homework significantly improved grades; (2) Students who failed in dormitory showed significant negative impact; (3) Students attributed outcomes mainly to external factors (57.3%).	(1) Low generalizability (single-institution sample); (2) Interaction effects untested; (3) Unstructured data unintegrated.
[6]	193 junior high students; Multiple linear regression	(1) Learning interest had the strongest effect (β=12.298); (2) High-frequency praise group scored significantly higher (57.3 vs 39.4 points).	(1) Limited sample representativeness; (2) Oversimplified teaching-method metrics; (3) Grade-level differences unanalyzed.
[3]	67 English majors; Regression analysis	(1) Significant positive correlation between English and math scores (r=0.702); (2) 1-point increase in English predicted 0.87-point math score rise; (3) No significant difference among same-teacher classes.	(1) Weak extrapolation (small sample); (2) Confounding variables uncontrolled; (3) Bias in manual threshold grouping.

5 Conclusion

This study systematically reveals the complex mechanism of action of factors influencing students' academic performance through machine learning methods, breaking through the theoretical limitations of traditional statistical analysis in dealing with high-dimensional nonlinear relationships and dynamic interaction effects. The core findings indicate that academic performance is jointly shaped by three dimensions: individual initiative, teaching management, and environmental regulation. Among them, the learner's own motivation level, self-efficacy, and learning strategy selection form the decisive basis, and their influence weight is significantly higher than that of teaching resource allocation and environmental support factors. The research further quantified the nonlinear synergy laws among key factors: positive learning motivation and self-efficacy form a positive reinforcement cycle, while excessive workload weakens the benefits of knowledge accumulation. Meanwhile, the analysis of subject differences reveals the prominent role of cognitive strategies in mathematics courses and the special moderating effect of psychological states on the performance of medical courses. These findings verify the core position of the "learner-centered" theory from a computational empirical perspective and provide a key basis for the construction of dynamic models of educational factors.

Looking to the future, theoretical exploration should be committed to building a more explanatory dynamic framework. The primary direction is to deepen the time evolution theory of the complex educational system, and to track the reorganization rules of factor weights during the transition of learning stages through time series modeling, such as the dynamic process of anxiety relief in the first year and the transformation of professional identity in the second year. Secondly, it is necessary to promote the integrated research of interdisciplinary mechanisms, combining the working memory theory of cognitive psychology, the emotion regulation mechanism of neuroscience and machine learning algorithms to deeply analyze the differentiated demands of different knowledge types (such as logical deduction in mathematics and image memory in medicine) for learning paths. Finally, a unified theory of multi-dimensional interaction should be developed, and a mathematical model that can integrate the internalization process of motivation, the design of curriculum correlation, and the effects of environmental interference should be established, with a focus on exploring the educational intervention significance of the critical point effect among factors.

The theoretical value of this research lies in transforming machine learning into a "computational laboratory" for exploring educational laws, promoting the academic community's paradigm shift from static attribution to dynamic mechanism analysis. By revealing the dialectical relationship between individual initiative and environmental constraints, as well as the shaping effect of disciplinary characteristics on the learning mechanism, a new path has been opened up for constructing a universal theory of complex educational systems. In the future, reinforcement learning can be further integrated to simulate the long-term intervention effects, enabling educational theories not only to explain reality but also to proactively optimize the learning ecosystem.

References

1. Cheng, W.: Research on the impact of English learning motivation on academic performance of vocational undergraduate students. University (5), 181–184 (2025)
2. Feng, L., Li, J., Li, Y., Yan, Q.: Analysis of the correlation and influencing factors of open education course grades: Based on a survey of 292 students majoring in pharmacy at Beijing Open University. Journal of Hainan Open University 24(1), 78–86 (2023)
3. Ji, H.: Analysis of factors affecting mathematics scores of English major students: An empirical study based on SPSS statistical analysis. Journal of Higher Correspondence Education (Natural Science Edition) 24(3), 67–68 (2011)
4. Zhou, X., Ma, C., Liu, Q.: Analysis of influencing factors of students' academic performance based on regression method. Science and Technology Innovation and Application 12(23), 103–106 (2022)
5. Bai, X., Xu, H., Fu, D.: Exploration of the influencing factors of academic performance of medical imaging students. Journal of Multimedia and Network Teaching in China (Upper Edition) (8), 184–187 (2023)
6. Zhao, H., Feng, B., Liu, K.: Analysis of the influencing factors of junior high school mathematics scores based on multiple linear regression. Journal of Longdong University 36(2), 131–138 (2025)
7. Pan, X., Guo, Q., Lin, N.: Investigation on the influencing factors of undergraduate students' advanced mathematics scores: An analysis based on logistic regression model. College Mathematics 37(4), 60–69 (2021)
8. Xiao, Q., Zhang, L., Shi, E.: Statistical analysis of influencing factors of college students' mathematics scores in different grades. Journal of Mathematical Education 24(4), 53–56 (2015)
9. Hu, B., Cui, Z.: Statistical analysis of influencing factors of advanced mathematics course academic performance in local colleges and universities: A case study of Chuzhou University. Journal of Xichang University (Natural Science Edition) 37(2), 116–126 (2023)
10. Chi, L., Xin, Z.: Measurement of college students' learning motivation and its relationship with self-efficacy. Psychological Development and Education (2), 64–70 (2006)