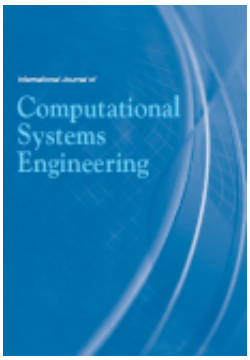
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Systems Engineering

ISSN online: 2046-3405 - ISSN print: 2046-3391
https://www.inderscience.com/ijcsyse

Analysis of factors affecting college students' academic performance based on linear regression

Ru Huang

DOI: 10.1504/IJCSYSE.2026.10077793

Article History:

Received:	13 August 2025
Last revised:	30 October 2025
Accepted:	22 January 2026
Published online:	29 April 2026

Analysis of factors affecting college students' academic performance based on linear regression

Ru Huang

Academic Affairs Office,
Guilin University of Aerospace Technology,
Jinji Road, Qixing District, Guilin, Guangxi, China
Email: wyy868707@163.com

Abstract: The article examines determinants of college students' academic performance (AP) by using an extended linear regression model (ELRM). Unlike traditional linear regression models (TLRMs), which fail to properly account for the time-varying effects and interactions of dynamic behavioural determinants, the ELRM incorporates time series characteristics to model the temporal and interactive effects on AP. Based on data spanning several semesters, the study finds that short-term learning habits (LHs), such as cramming, have a greater immediate contribution to grades compared to long-term studying habits. The ELRM performs better than conventional models with a considerably lower mean squared error (MSE = 0.81) and a better coefficient of determination (R^2 = 0.985), demonstrating its enhanced predictive ability and stability across semesters. The study highlights the value of temporal determinants and interaction effects in explaining the dynamics of AP, providing recommendations for enhancing educational interventions and policy.

Keywords: academic performance; extended linear regression; study habits; time series analysis; learning behaviour; cramming; dynamic interaction; data mining.

Reference to this paper should be made as follows: Huang, R. (2026) 'Analysis of factors affecting college students' academic performance based on linear regression', *Int. J. Computational Systems Engineering*, Vol. 10, No. 8, pp.1–13.

Biographical notes: Ru Huang graduated with a Master's in Education Economics and Management from the Guangxi Normal University and she obtained her Master's in Management in 2017. Currently, she works at the Academic Affairs Office of Guilin Aerospace Industry College as an Assistant Research Fellow. Her main research focus on education management and higher education.

1 Introduction

The academic performance (AP) of students in college has been an important area of educational research for a long time. An understanding of the determinants of AP is important not just for the academic achievement of students but also for enhancing teaching methods, curriculum development, and policy-making. AP is affected by a number of variables ranging from internal attributes of students such as learning habits (LHs) of students, motivation, and time management to external factors such as teaching quality, peer interaction, and environmental support systems. The interaction among these variables presents a complex dynamic that cannot be explained by conventional models (Al Husaini and Shukor, 2022). Traditionally, linear regression models (LRMs) have been employed to analyse the correlation between the learning behaviour (LB) of students and their AP. LRMs provide an easy approach to identifying key variables and determining their effect on AP (Alani and Hawas, 2021). However, linear regression has limitations in handling the temporal variation of LB and its interactive effects across time. In particular, factors like short-term learning activities, (e.g., cramming) and their varying degree of correlation with long-term study habits are often overlooked. This limitation is especially evident in analysing data spanning multiple semesters or academic years, where the impact of particular behaviours might differ across time.

Recent advancements in statistical modelling and ML offer promising approaches to remedy these limitations. Methodologies such as time series analysis, panel data regression, and dynamic modelling have been posited to capture the complex temporal dependencies and interaction effects between LBs and AP. Specifically, time series models allow the investigation of the ways in which students' behaviours evolve over time and how such changes relate to their performance outcomes (Hussain et al., 2021). However, despite these improvements, existing dynamic models are still limited in their accurate portrayal of the interaction effects of study habits and temporary learning input, especially when considering the nonlinear nature of these behaviours (Bell, 2025). This study proposes an extended linear regression model (ELRM) that incorporates time series analysis to address the dynamic and interactive factors influencing AP. By incorporating lagged

variables and interaction terms, the ELRM can capture the temporal changes in the relations between LBs and academic achievement. This model not only improves the predictive power of AP but also provides deeper insights into the complex mechanisms underlying students' academic success. The suggested model aims to address the shortcomings in conventional LRMs by including both temporal dynamics and the interaction between short-term and long-term LBs, which are necessary to understand how students' AP changes over various time periods.

The main goal of this paper is to create a stronger, dynamic model that will more accurately predict AP and describe the process by which various LHs and short-term learning investments evolve and interact over time. The next several sections describe the development of the ELRM, data collection, and methodology for building the model. This introduction provides the backdrop for an in-depth examination of how dynamic behaviour, like cramming, interacts with long-term study habits and what these interactions imply for enhancing educational outcomes.

1.1 Motivation and research gap

The incentive for the current study arises from the complexity of learning environments and the demand for more precise prediction models. As the LB of students becomes more heterogeneous and intricate, so does the demand for models that have the capacity to incorporate time-varying effects and the synergies among factors. Conventional LRMs, although helpful, do not capture the dynamic behaviour of students' behaviour, especially in modelling how behaviour unfolds over time and how the behaviours interact with each other in affecting AP. This is particularly glaring when considering behaviour like cramming, which affects AP in the short term but interacts with long-term study habits. Recent studies have attempted to integrate time series and dynamic features into predictive models of AP. For example, researchers have employed ML approaches like long short-term memory (LSTM) networks and advanced regression models to develop a better understanding of temporal dependencies. However, these models often fail to include important interaction terms or sufficiently model the nonlinear relationships between various LBs and academic success. While time series models provide valuable insights, there remains a gap in their capability to capture the full complexity involved in the interplay between long-term study habits and short-term learning activities.

By incorporating lagged features and interaction terms, this research aims to address this gap by presenting a model that not only models the temporal dynamics of the LB of students but also explains the interactive effects between behaviours such as consistent study habits and short-term cramming. The suggested ELRM presents a more unified approach that can greatly enhance the accuracy of AP predictions as well as our understanding of the underlying drivers of academic achievement.

1.2 The organisation of the manuscript

The organisation of the paper is as follows: Section 2 reviews the literature on the subject, covering past attempts at modelling and predicting AP through traditional regression models as well as newer ML approaches. Section 3 introduces the new ELRM, covering mathematical formulation, methodology, and data pre-processing approaches taken. Section 4 covers experimental setup and data collection, with discussion of the results covered in Section 5. Section 6 concludes, summarising the main findings and implications for educational practice. The current research aims to contribute to the body of literature in educational data mining by formulating the ELRM as a more effective tool for clarifying the complex influences on academic success. By covering the complicated relationships between LBs and temporary learning inputs, the model derives new insights into the dynamics of these variables across time, thus enabling the deployment of more accurate and effective teaching interventions.

2 Related work

Numerous investigations have started to concentrate on how to improve the prediction accuracy of students' AP through different algorithms and models, especially through dynamic data analysis to overcome the limitations of traditional static models. These studies have gradually developed more multidimensional and comprehensive prediction frameworks, especially in terms of the time dependency and interaction effects of behavioural mechanisms. For example, Al-Ali et al. (2024) analysed the impact of students' behaviour and preferences on AP by using methods such as K-means clustering, LSTM neural network and principal component analysis, and found that AP is significantly affected by gender, department affiliation and certificate course participation, but performs poorly in predicting minority classes. Concurrently, Tao (2025) employed the advanced Penguin search optimised adaptive boosting (APSO-AdaBoost) machine learning (ML) model to examine data of various student groups and suggested that the model is capable of effectively discovering key performance aspects, particularly when working with students of different cultures, and it can also offer a sound foundation for the support of the educational strategy optimisation. In these ways, scientists have reached the opinion that they are able to make correct predictions in the situation of diversified student groups. Additionally, Kukkar et al. (2024) suggested a new way for forecasting student academic success by the combination of ML models, such as recurrent neural networks, LSTM networks, and random forests (RFs). This method can precisely grasp the time changes in student data. Liang et al. (2024) utilised five ML models to forecast the AP of online course students and revealed that the gradient boosting regression model gave the best results in terms of prediction accuracy. They also pointed out that intellectual education scores, the number of homework completed, and the live broadcast

viewing rate and playback viewing rate in online LB are the key factors affecting students' final exam scores. This study also focuses on the impact of real-time factors in LB on grades, further promoting the development of behavioural analysis models.

In another study, Qureshi et al. (2023) employed a structural equation model to examine the influence of social factors on collaborative learning and student engagement and found that social factors such as peer and teacher interaction, social presence, and social media use can promote collaborative learning and student engagement, thereby improving students' learning performance. Furthermore, Feraco et al. (2023) utilised Bayesian path analysis to investigate how soft skills, outside interests, emotions related to success, self-directed learning, drive, and mental aptitude affect students' scholastic performance and contentment with life. Mahmud et al. (2022) analysed the factors affecting students' AP through multiple linear regression. The results showed that students' hometowns and pre-class preparation time had a significant impact on AP, especially for students in rural areas, who demonstrated better AP. In contrast, students with longer preparation times had lower cumulative grade point average (CGPA). Other factors did not show a significant effect. Shou et al. (2024) put forward a model for forecasting student success using multi-faceted time series data analysis. Combining students' LB, assessment scores, and demographic information can effectively identify risky students and achieve personalised education. Although these studies have achieved remarkable results in predicting AP, there are still challenges in further improving and expanding LRMs, strengthening interaction effect analysis, and improving the prediction ability of minority classes.

Over the past few years, numerous investigations have examined the function of LB, motivation, and teaching models in academic achievement. For example, Tagud and Valle (2023) adopted descriptive correlation design and causal design, and used multiple linear regression to analyse students' LHs, academic pressure, and concentration. They found that academic pressure had a negative impact on AP, while concentration had a positive impact, indicating that LHs play an important role in AP. Nieberding and Heckler (2021) used latent class analysis to study the relationship between students' procrastination behaviour and AP, American College Test (ACT) scores, and gender, and found that students who procrastinate have lower grades, and the procrastination time is closely related to the non-exam scores.

Typically, women exhibit a lower tendency to put things off compared to men, and procrastination behaviour plays a mediating role in gender differences. In previous studies, many analyses of AP have overlooked the time dependence of behavioural habits and their interactive effects. Xiong et al. (2021) longitudinal study effectively filled this gap, exploring the bidirectional relationship between parental involvement and AP and the mediating role of adolescent academic participation. It was found that AP had a positive lagged effect on parental involvement, and this effect was fully mediated by academic behaviour participation in boys, while girls did not show a significant mediating effect. Crowther and Briant (2021) longitudinal study employed regression analysis to assess the connection between college entrance grades, 1st-year performance, and academic success. To further reveal the differences in teaching modes, Salem et al. (2024) compared the effects of in-person, online and hybrid learning modes on the performance of business management students through longitudinal data analysis and found that the learning motivation of the online learning mode significantly improved students' learning satisfaction and had a greater impact on AP, while the hybrid learning mode.

However, improving learning satisfaction had a relatively low impact on AP. Ho et al. (2021) examined the elements influencing undergraduate student contentment in emergency remote education at a private university in Hong Kong, utilising regression analysis and a comparison of ML models. Finally, Pepe and McCollum (2023) compared the scores of students in remote synchronous teaching and in-person teaching. The work found that the GPA of students in remote teaching at the end of the semester was significantly lower than that of students in face-to-face teaching, further verifying the impact of different learning modes on AP. Existing research focuses on the influence of LB, motivation, and teaching mode. Still, it often fails to explore their interactive effects and time-dependent mechanisms in depth, especially in terms of dynamics and lag effects.

3 Algorithm design

3.1 Data pre-processing and feature extraction

Data pre-processing cleans AP data by dealing with missing values through mean interpolation, eliminating outliers by Z-score standardisation, and removing duplicates. After consistency assurance, the data is standardised by Z-score normalisation to deal with dimensional differences among variables. In feature extraction, pertinent variables like weekly study time, review habits, class attendance, and extracurricular activities are extracted. These variables also cover lagged features to capture the temporal effect of LBs on AP. For example, study time includes data for the current week as well as the preceding two weeks to account for temporal trends. Short-term learning inputs, (e.g., cramming) and interaction features are also included to capture the synergy between LHs and instantaneous input. The generalisation capability of the model is improved by correlation analysis and stepwise regression, which remove redundant features and deal with multicollinearity based on the variance inflation factor (VIF). The final feature set covers lag features, temporal changes, peak learning inputs, and interaction effects (Kyriazos and Poga, 2023). The process is shown in Figure 1.

3.2 Dynamic impact mechanism modelling based on linear regression

In constructing the ELRM, this study first incorporates time series characteristics within the traditional linear regression framework, capturing the dynamic interaction effects between AP and LB by applying lag terms and interaction terms (Yildiz Durak, 2025). To ensure that the model can handle time-varying and interactive characteristics, this study designs multiple lagged features and interaction terms to reveal the long-term impact and short-term fluctuations of LB on AP. Basic model framework. The constructed ELRM can be expressed as:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{t-i} + \sum_{j=1}^{q} \gamma_j Z_{t-j} + \epsilon_t \tag{1}$$

where Y_t is the AP at time t; X_{t-i} is the LB variable at the i^{th} lag period; Z_{t-j} is the LH variable at the j^{th} lag period; and ϵ_t is the error term. Application of lagged terms: To capture the time-varying relationship between LB and AP, this study applies lagged terms for each LB variable and AP. Assuming that the impact of a student's weekly study time X_t on AP depends not only on the current moment but also on several previous moments. Therefore, this study adds lagged variables to the model:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} \tag{2}$$

In which ϕ_i represents the RC of the lag term, and p is the lag period. This enables the model to effectively capture the historical effect of LB. Construction of interaction terms: to reveal the interaction effect between LB and temporary learning investment, interaction terms are also applied to the model. The selection of lag order p and interaction term dimension q is based on a trade-off between prediction stability and model simplicity under cross-validation. An excessively small p fails to capture persistent behavioural effects, while an overly large p introduces redundant noise and exacerbates multicollinearity. Similarly, insufficient interaction terms may overlook critical synergistic mechanisms, whereas excessive terms lead to overfitting. Therefore, this study employs grid search to explore parameter combinations within a predefined range (p, $q \in \{1, 2, 3, 4\}$), aiming to minimise mean squared error (MSE) through 10-fold cross-validation while monitoring VIF and residual autocorrelation. This approach ensures that selected parameters enhance predictive performance while maintaining structural stability of the model. The construction method of the interaction term is:

$$X_t \times Z_t = \sum_{i=1}^{p} \sum_{j=1}^{q} \theta_{ij} X_{t-i} Z_{t-j} \tag{3}$$

where θ_{ij} is the RC of the interaction term; X_t is the learning time; Z_t is the review frequency. ELRM optimisation: to further enhance the model's predictive ability, this study optimises it by adjusting the number of lag terms p and the number of interaction terms q. The optimal combination of lag terms and interaction terms is selected through cross-validation to minimise the mean square error:

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - \hat{Y}_t \right)^2 \tag{4}$$

In this process, this study employs the stepwise regression method and the ridge regression method to gradually eliminate redundant variables, controlling model complexity through regularisation to prevent overfitting (Amdee, 2024). Enhancement of time series features: to fully utilise time series data, this study not only applies lag features and interaction terms but also considers the autocorrelation of the model. To this end, this study uses an autoregressive model to model the residuals:

$$\epsilon_t = \sum_{k=1}^{r} \phi_k \epsilon_{t-k} + v_t \tag{5}$$

where v_t represents white noise, and r denotes the order of the autoregressive process. The autoregressive model helps capture the time dependence in the model error; thereby improving the model's fitting accuracy. Fitting of the ELRM: the ELRM is customised by reducing the subsequent objective function.

$$\min_{\beta,\gamma} \sum_{t=1}^{n} \left(Y_t - \beta_0 - \sum_{i=1}^{p} \beta_i X_{t-i} - \sum_{j=1}^{q} \gamma_j Z_{t-j} - \sum_{i=1}^{p} \sum_{j=1}^{q} \theta_{ij} X_{t-i} Z_{t-j} \right) \tag{6}$$

The RCs are estimated using the least squares method to obtain the optimal parameters of the model. Stability and time dependence of the model: To test the stability of the model, this study employs a recursive algorithm to assess the model's consistency over time by comparing the RCs across different time intervals. At the same time, the long-term and short-term effects of LB on grades are detected by volatility analysis:

$$\Delta Y_t = \sum_{i=1}^{p} \beta_i \Delta X_{t-i} + \sum_{j=1}^{q} \gamma_j \Delta Z_{t-j} \tag{7}$$

where ΔX_{t-i} and ΔZ_{t-j} are the changes in AP and LB variables, respectively. Stability testing helps ensure the model's adaptability to changes in the data. Final model formula: after parameter optimisation, interaction term construction, and time series feature enhancement, the final form of the model is:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{t-i} + \sum_{j=1}^{q} \gamma_j Z_{t-j} + \sum_{i=1}^{p} \sum_{j=1}^{q} \theta_{ij} X_{t-i} Z_{t-j} + \epsilon_t \tag{8}$$

3.3 Model training and optimisation

In the training and optimisation of the model, to improve prediction performance and remedy overfitting and

underfitting, this research tunes hyperparameters and chooses the optimal model. For training, a cross-validation strategy is used, along with regularisation and hyperparameter tuning for better generalisation. To provide robustness to heterogeneous datasets, stratified sampling splits the data into training and evaluation sets, and k-fold cross-validation is applied to reduce estimation bias by reshuffling the data in a systematic manner (Bates et al., 2024). Both the training loss and validation prediction error are monitored in each iteration to manage model complexity and avoid overfitting, making high accuracy on unseen samples possible.

$$\mathrm{CV} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_{ij}\right)^2\right) \tag{9}$$

In which, y_{ij} and $\hat{y}_{ij}$ are the actual and predicted values of the i^{th} fold, respectively, and n_i is the number of samples in the i^{th} fold. There is a significant correlation between model performance and hyperparameter configuration, among which the lag parameter p and the interaction term dimension q constitute the key regulatory factors. In order to determine the optimal parameter combination, this study constructs a multidimensional parameter search space and adopts a systematic parameter optimisation strategy. By calculating the cross-validation loss function value under different parameter configurations, the parameter combination that minimises the error of the validation set is finally selected as the optimal solution. The objective function of the grid search is:

$$\hat{\theta} = \arg\min_{\theta}\sum_{i=1}^{k}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_{ij}\right)^2\right) \tag{10}$$

where θ is the set of hyperparameters to be tuned. Grid search searches for the optimal solution by traversing all possible combinations in the preset hyperparameter space. To prevent overfitting, this study applies regularisation technology to the LRM, utilising Lasso regression (L1 regularisation) and Ridge regression (L2 regularisation) to control model complexity (Mohamed et al., 2025). The objective function of Lasso regression is:

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right) \tag{11}$$

In which, λ denotes the parameter that regulates the penalty amount and hence, the degree of regularisation.

Ridge regression's objective function is:

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right) \tag{12}$$

In this research, the extent of regularisation is regulated by changing the value of λ so that the model is not overfit. The training process of the model in this study involves selecting the best-performing model by minimising the square errors and the root mean square error (RMSE) of the training. MSE and RMSE signify how good the model's prediction is. RMSE is the square root of MSE, and the expression is:

$$RMSE = \sqrt{MSE} \tag{13}$$

RMSE gives the real size of the mistake, which makes it easy to get an intuitive insight into the prediction mistake of the model. The leave-one-out cross validation (LOOCV) technique (Geroldinger et al., 2023) is one option for checking the model's generalisation potential. The error computation formula of LOOCV is:

$$\mathrm{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{14}$$

where y_i denotes the observed value, and $\hat{y}_i$ signifies the estimated value obtained by excluding the i^{th} sample. The residuals are also being studied in this research to confirm the stability of the model training and check the distribution of prediction errors in various time intervals. Residuals' autocorrelation is investigated in this research to verify that the model does not have any time dependence. The expression for the residual autocorrelation is:

$$\mathrm{ACF}(\tau) = \frac{\sum_{t=1}^{n-\tau}\left(y_t-\hat{y}_t\right)\left(y_{t+\tau}-\hat{y}_{t+r}\right)}{\sum_{t=1}^{n}\left(y_t-\bar{y}\right)^2} \tag{15}$$

According to equation (15), τ refers to the time period and $\bar{y}$ is the average of the sample. The model's stability is also tested by analysing the residuals' autocorrelation to verify that no time series characteristics have been missed. To make the training process as efficient as possible, this study has changed several important hyperparameters with the grid search method and used cross-validation to check the model's performance. Table 1 provides the detailed configuration and the results of the evaluation.

Table 1 illustrates the grid search parameters and the corresponding evaluation scores for various hyperparameters that were used during the training and optimisation of the model. It indicates the range of values and the way of selecting hyperparameters such as the number of lags, the number of interaction terms, Lasso and Ridge regularisation parameters, and cross-validation folds. Through adjusting these hyperparameters and verifying the model's performance in different configurations, the most suitable combination of hyperparameters is eventually chosen to enhance the model's forecasting ability and trustworthiness.

Figure 1 Data pre-processing and feature extraction flowchart (see online version for colours)

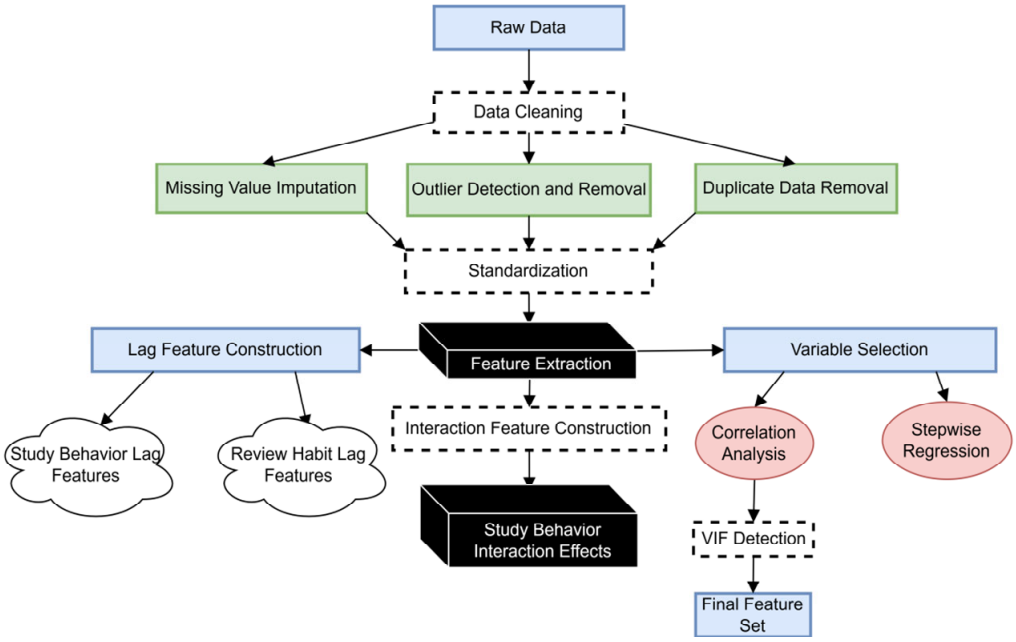
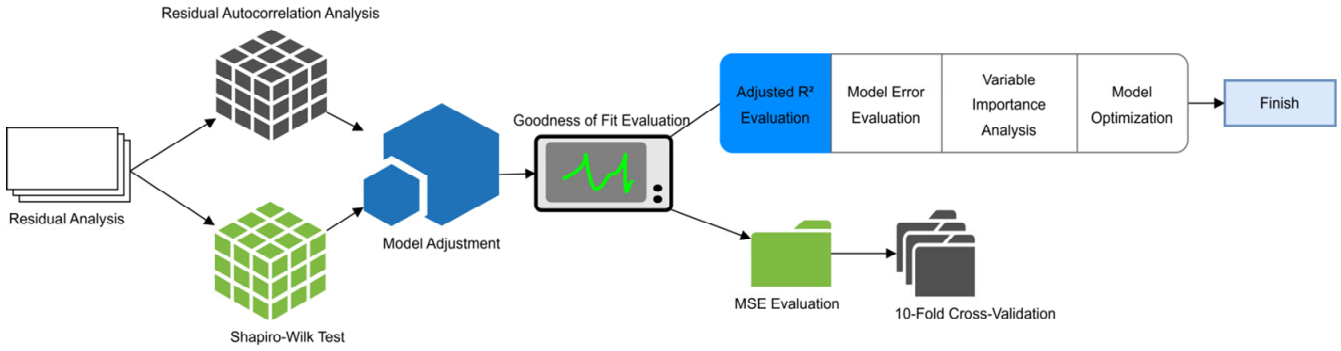


Table 1 Hyperparameter configurations and assessment outcomes during model instruction and enhancement

Hyperparameter/setting	*Range of values*	*Selection method*	*Performance evaluation metrics*	*Evaluation result*
Lag period (p)	1, 2, 3, 4	Grid search	MSE, RMSE	Optimal lag period: 2
Number of interaction terms (q)	1, 2, 3, 4	Grid search	MSE, RMSE	Optimal interaction terms: 3
Lasso regularisation parameter (λ_1)	0.001, 0.01, 0.1	Grid search	MSE, RMSE	Optimal λ_1: 0.01
Ridge regularisation parameter (λ_2)	0.001, 0.01, 0.1	Grid search	MSE, RMSE	Optimal λ_2: 0.01
Cross-validation folds (k)	5, 10	Fixed as 10-fold cross-validation	CV error	CV error: 0.023

Figure 2 Model verification and diagnosis (see online version for colours)



3.4 Model validation and diagnosis

Following model estimation, validation is required to check statistical adequacy and model parameters. Residual diagnostic analysis diagnoses model fit through the examination of residuals. When residuals exhibit random noise with no appreciable autocorrelation, the model has succeeded in capturing linear relationships. Patterned residual plots indicate model problems, including omitted variables or unmodelled nonlinearities. Residual normality is examined with the Shapiro-Wilk test, and if normality is not met, the model can be modified or data can be transformed. Residual autocorrelation is examined for time-dependency, calling for model modification or the inclusion of more time lags.

Goodness-of-fit statistics such as R^2, adjusted R^2, and MSE measure predictive accuracy. R^2 provides explanatory power, where values nearer to 1 indicate a good fit, and adjusted R^2 considers overfitting, while MSE measures errors in predictions. A time stability test looks for performance consistency across time, and the presence of significant error variation indicates that the model is

time-dependent. Cross-validation, and specifically the 10-fold version, is a way to assess the model's generalisability while reducing the chance of overfitting. Lasso regression is used in the analysis of variable importance to identify key factors influencing AP, to optimise the model, and enhance its interpretative simplicity. The steps are summarised in Figure 2.

4 Experiment and verification

4.1 Experimental design

In this experiment design, student grade data and LB data are gathered from several colleges and universities so that each student's data cover at least one semester to reflect dynamic behaviour. Sources of data include grade reports, self-assessment questionnaires, and LB tracking platforms. Weekly study hours, frequency of review, class attendance, and extracurricular activity variables are tracked. The data cover several semesters or academic years so that long-term trends in AP can be analysed. Time series data are recorded weekly to dynamically reflect AP changes, and the data cover students from at least three universities with different academic levels and teaching models to guarantee diversity.

Time series corresponding to different LBs, including extracurricular learning time and class participation, are employed to characterise students' learning input and approaches. In order to model the time-lagged influence of behaviours on AP, lag terms of every LB are included. Data cleaning includes filling in missing values with the mean-based approach, outlier detection using the Z-score approach, and removing abnormal data. Following data standardisation to remove scale discrepancies, stationarity of time series is achieved through applying the difference approach. Feature engineering is conducted to distil pertinent lagged features, which gives a more precise characterisation of students' LBs.

4.2 Evaluation of the impact of study habits on grades

Variables related to LHs are extracted, mainly including learning frequency, review frequency, class participation, and extracurricular activities. These variables are modelled through time series analysis, taking into account the lag effect and cyclical changes of each LB. The ELRM is employed to examine the long-term relationship between LHs and AP, with a focus on assessing the significance of LHs in the model, particularly the impact of various LHs on AP. The relevant outcomes are demonstrated in Figure 3.

Figure 3 illustrates the dynamics of AP and study habits of five students over four semesters. The grade of Student 1 decreases from 75 in Semester 1 to 60 in Semester 3, whereas the score of Student 2 wildly swings from 62 in Semester 2 to 99 in Semester 3. Review frequency is different across students, with Student 3 reviewing 8 times/week in Semester 3, whereas Student 5 reviews 5 times/week. Participation in class and extracurricular activities also vary across students and semesters. These data indicate a complicated relationship between AP and study habits, with study and review frequency clearly affecting grades. The ELRM embeds time series characteristics to describe the dynamics and interaction effects of LBs, providing an understanding of how different learning approaches affect AP changes over semesters. The results are encapsulated in Table 2.

Table 2 RCs and significance test results of LH variables

Learning habit variable	*Standard error*	*Coefficient*	*t-value*	*p-value*
Review frequency	0.08	0.42	5.25	< 0.001
Study frequency	0.07	0.31	4.43	< 0.001
Class participation	0.05	0.15	3	0.003
Extracurricular activities	0.04	–0.1	–2.5	0.013

Table 2 shows the regression coefficients (RCs), standard errors, t-values, and p-values of the four LH variables in the ELRM model. Review frequency has the largest positive impact on AP, with an RC of 0.42 and a p-value less than 0.001, thus highlighting its significant contribution. Learning frequency also has a significant impact on AP, with an RC of 0.31 and a p-value less than 0.001. Class participation has a smaller but still significant effect, as shown by an RC of 0.15 and a p-value of 0.003. Extracurricular activities have a negative impact, as reflected by an RC of –0.1 and a p-value of 0.013, which suggests that overindulgence in extracurricular activities may detract from academic focus. The model highlights the fundamental roles of systematic review and regular learning frequency in improving AP, while class participation and extracurricular activities have more modest effects.

4.3 Evaluation of the impact of temporary learning input on AP

Aiming at the dynamic influencing mechanism of temporary learning input on college students' AP, this study collects data on individualised LBs over multiple semesters to conduct an empirical analysis. Figure 4 shows the specific observation values of five students in four semesters, including the AP scores at the end of each semester and the corresponding records of concentrated study time before the exam, showing the corresponding relationship between students' actual performance and temporary learning input in different semesters.

Figure 4 demonstrates the relationship between AP and temporary study attempts, (e.g., last-minute studying) over several semesters. For example, in the first semester, Student 1 scores 75 after studying for 37 hours, while Student 2 scores 98 with 33 hours of studying. In later semesters, study time is different. While temporary study attempts show a positive relationship with academic scores,

their efficacy is short-term in nature, particularly as examinations draw near. Overreliance on short-term study approaches can be disruptive to consistent learning patterns, indicating that cramming may improve immediate grades but not necessarily long-term academic achievement. The impact of temporary study investment at different intervals before exams is shown in Table 3.

Table 3 Analysis of the lag effect of temporary learning input

Lag period (weeks)	*Regression coefficient*	*95% confidence interval*	*p-value*
One week before exam	0.38	[0.32, 0.44]	< 0.001
Two weeks before exam	0.18	[0.10, 0.26]	0.002
Three weeks before exam	0.05	[–0.03, 0.13]	0.21
Four weeks before exam	–0.02	[–0.10, 0.06]	0.621

Table 3 shows the RCs and statistical significance of learning input and AP at varying lag periods prior to the exam (1 to 4 weeks). One week prior to the exam, learning input has the largest positive effect on grades (RC = 0.38, $p < 0.001$). This effect is still significant two weeks prior to the exam (RC = 0.18, $p = 0.002$) but diminishes over longer durations. The RCs for 3 and 4 weeks prior to the exam are 0.05 and –0.02, respectively, with corresponding p-values of 0.21 and 0.621, reflecting no significant effect. This indicates that last-minute studying contributes more to grades, in accordance with the recency effect, where intensive learning proximal to the exam serves to enhance performance. Early learning investments, however, demonstrate diminishing returns, where the effect becomes weaker as time goes by, likely due to the decline in motivation or lack of effect from earlier study sessions. The lack of effectiveness of early learning investment, particularly 4 weeks prior to the exam, demonstrates the tendency for unconsolidated knowledge to deteriorate over time, emphasising the role of timing when resorting to cramming.

4.4 The impact of the interactive mechanism between LHs and temporary investment on long-term performance

Based on the need to analyse the interaction terms in the ELRM, this study constructs a time-lagged cross-feature matrix of the LH variable and the temporary input variable, forming the basis for calculating the interaction effect. This matrix covers the four core LHs in four semesters to ensure that the time-varying characteristics of behavioural synergy are captured. To intuitively demonstrate the existence and directional characteristics of synergy, this section generates a cross-semester interaction effect heat map, revealing the potential pattern of behavioural coupling. The outcomes are displayed in Figure 5.

Figure 5 presents the interaction effect between study habits and temporary input with coefficients of review frequency, study frequency, class participation, and extracurricular activities over semesters. In Semester 1, the interaction effect between review frequency and temporary input is positively 0.11, whereas extracurricular activities have a negative interaction of –0.43. The findings indicate that the influence of study habits on AP is highly related to temporary input. For example, the positive impact between review frequency and temporary investment indicates that increased review time prior to exams enhances grades, whereas too many extracurricular activities might decrease temporary learning input associated with AP. This dynamic interaction demonstrates the need to balance long-term study habits with short-term learning investment in influencing AP, indicating that temporary learning input greatly interacts with other LBs in the short-term pattern.

4.5 Comparison between the ELRM and the traditional linear regression model (TLRM) on different evaluation indicators

In this research, a five-dimensional appraisal system is proposed to investigate the superiority of the ELRM over the TLRM in identifying interaction effects. The TLRM is used as the reference. Important performance metrics, such as predictive accuracy and model complexity, are gleaned via synchronous training on the same data. All metrics are normalised, and cross-validation is employed to mitigate random fluctuations. This section contrasts the capacity of the two models for detecting interaction effects, with findings presented in Figure 6.

Figure 6 displays a comparative evaluation of the performance measures between ELRM and TLRM on five basic indicators: MSE, R^2, model complexity, interaction effect capture, and prediction accuracy. ELRM shows better performance compared to the conventional model, which is indicated by both a lower MSE (0.81 vs. 0.91) and a higher R^2 (0.985 vs. 0.417), reflecting improved accuracy and model fit. In addition, ELRM outperforms TLRM on model complexity (0.92 vs. 0.75), ability to capture interaction effects (0.88 vs. 0.75), and prediction accuracy (0.75 vs. 0.65). This superiority can be explained by ELRM's ability to incorporate time series features, thus capturing the dynamic interplay between LBs and AP, specifically the short-term consequences of 'cramming' in addition to long-term study habits. In contrast, TLRM assumes static relationships, which limits its ability to adequately represent these dynamics. ELRM not only improves prediction accuracy but also enhances model interpretability, reflecting higher sensitivity to changes in students' learning patterns while offering higher flexibility in forecasting AP.

Figure 3 The relationship between study habits and AP (see online version for colours)

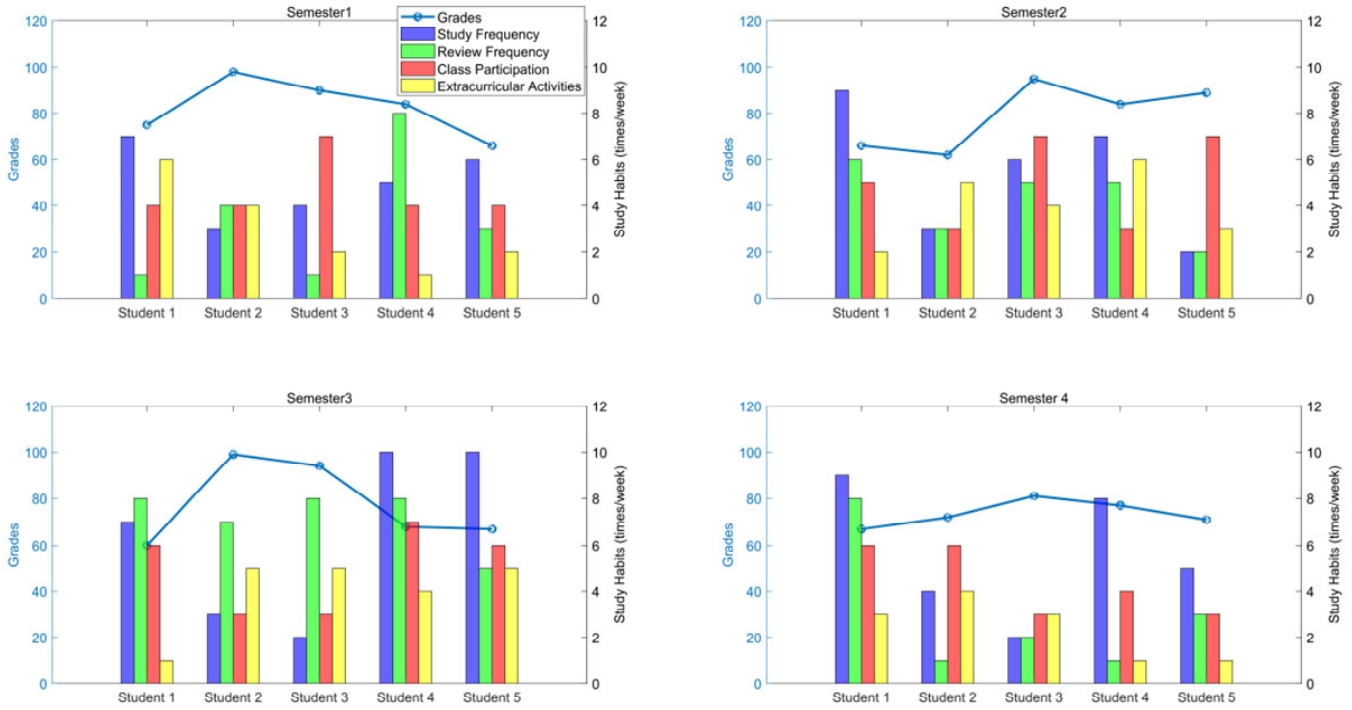


Figure 4 The impact of temporary learning investment on AP (see online version for colours)

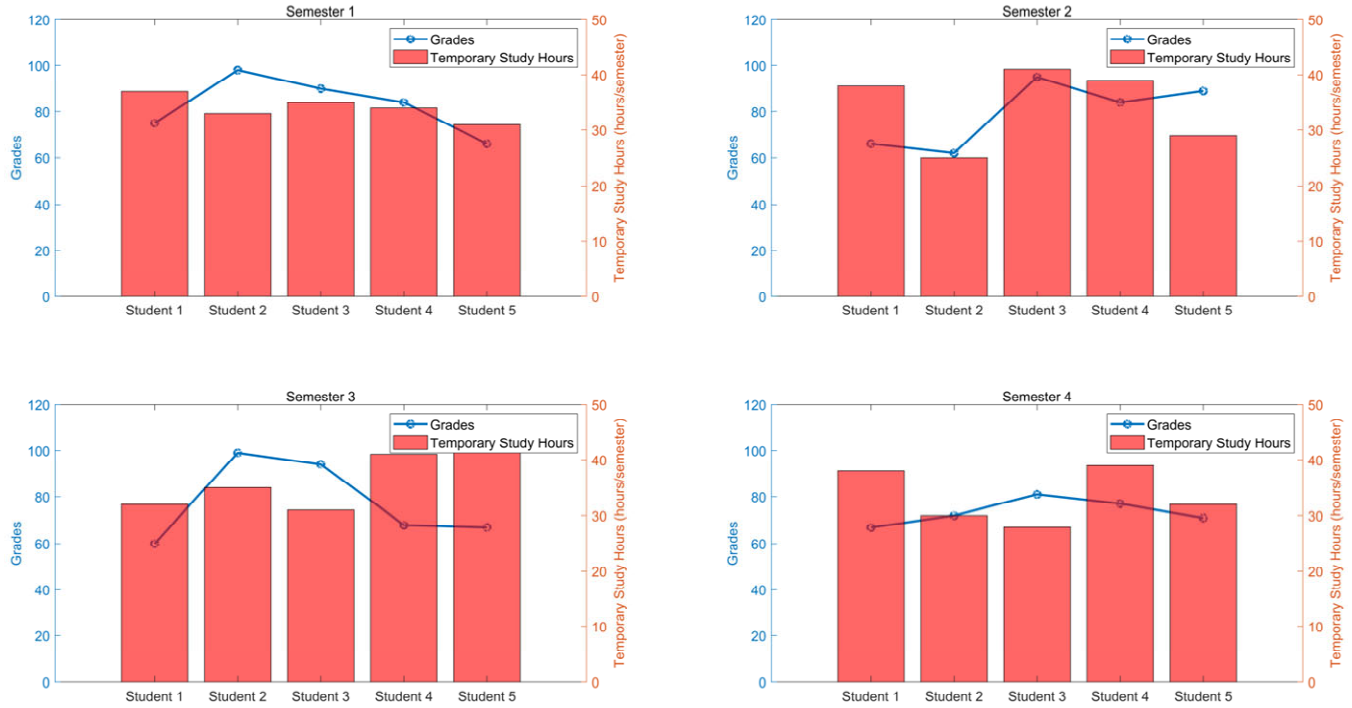


Table 4 Cross-group performance evaluation of ELRM on diverse datasets

Group category	*Subgroup*	*Sample size*	R^2	*MSE*	*RMSE*	*Interaction stability coefficient (ISC)*
Institution type	'Double first-class'	1,520	0.978	0.89	0.943	0.91
	Regular undergraduate	2,105	0.981	0.83	0.911	0.89
	Local applied-oriented	1,237	0.972	0.96	0.98	0.85
Discipline	STEM	2,340	0.983	0.8	0.894	0.9
	Humanities and Social Sci.	1,682	0.976	0.92	0.959	0.87

Table 4 Cross-group performance evaluation of ELRM on diverse datasets (continued)

Group category	*Subgroup*	*Sample size*	R^2	*MSE*	*RMSE*	*Interaction stability coefficient (ISC)*
Discipline	Business and Arts	840	0.97	1.01	1.005	0.83
Academic year	Freshman	1,210	0.965	1.08	1.039	0.82
	Sophomore-Junior	2,870	0.986	0.77	0.877	0.93
	Senior	782	0.968	1.03	1.015	0.84

Figure 5 The interactive effect of LHs and temporary investment (see online version for colours)

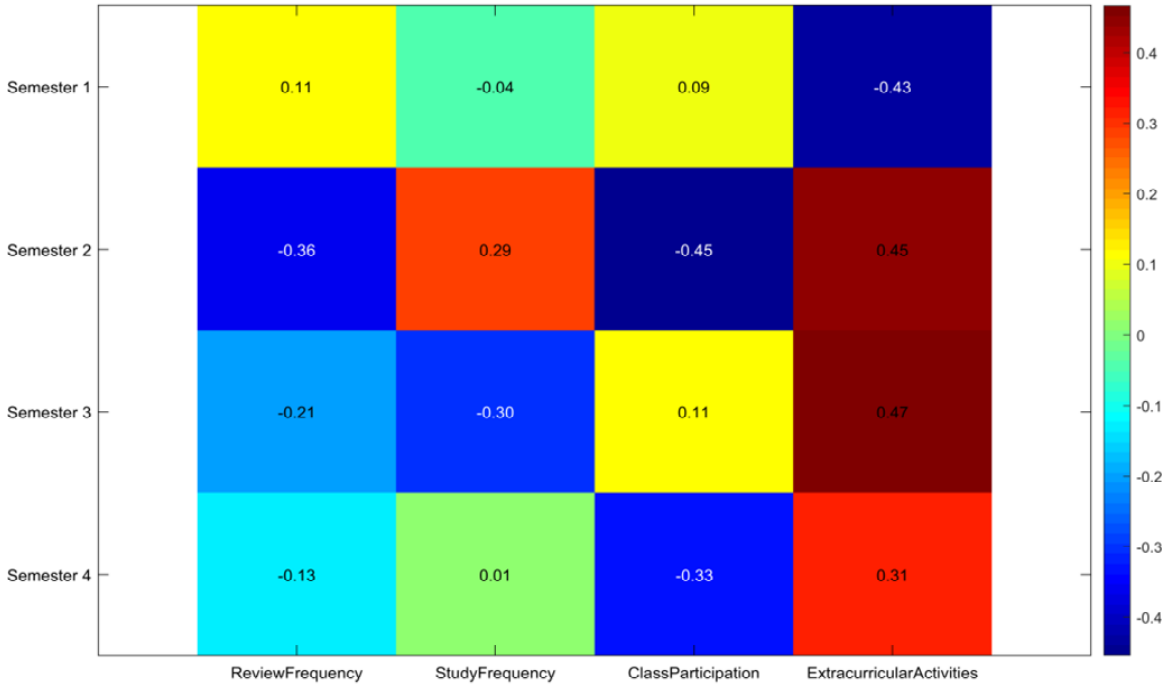


Figure 6 Radar chart comparing model performance (see online version for colours)

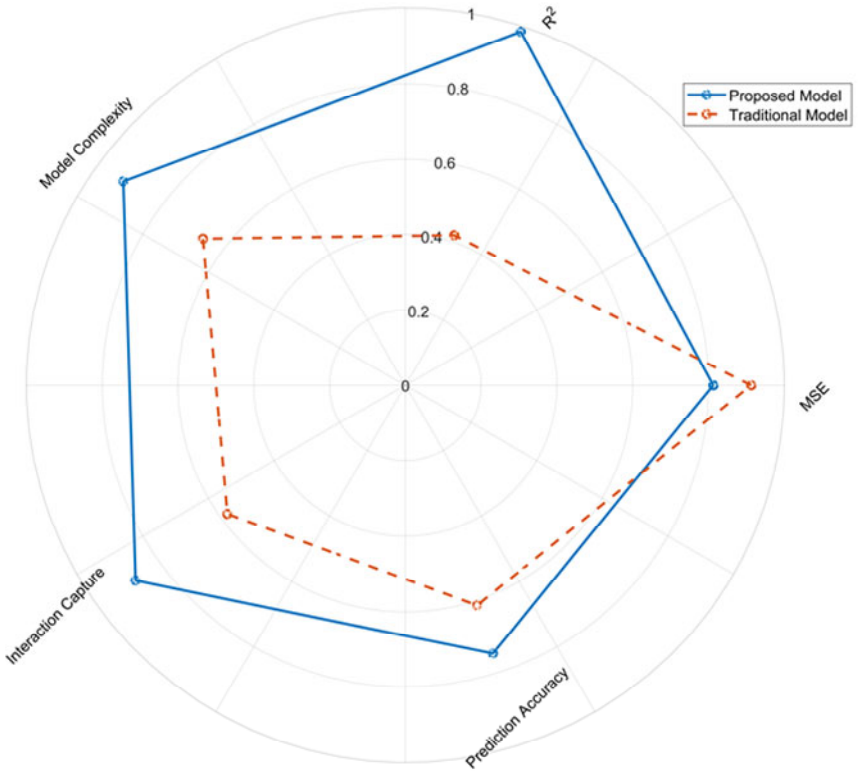


Figure 7 Comparison of the stability of the interaction effects of factors affecting AP across semesters (see online version for colours)

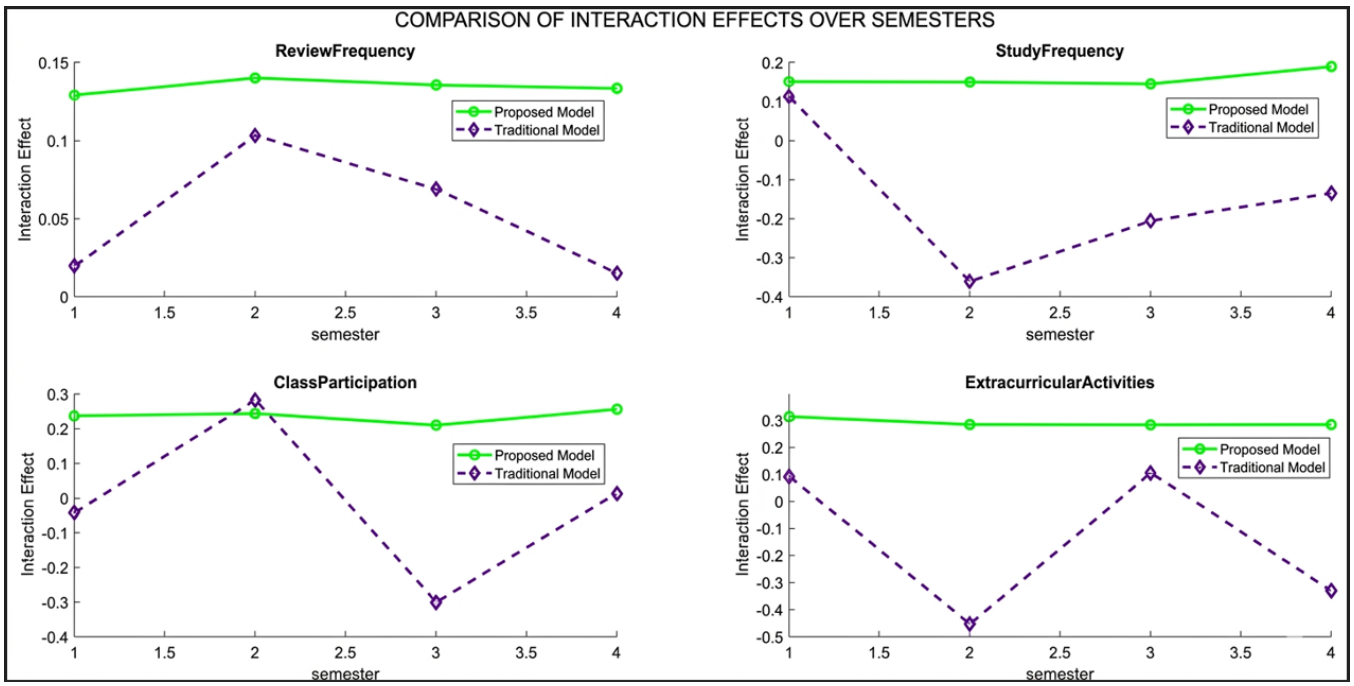


Table 5 Performance comparison of ELRM and advanced nonlinear models under unified experimental conditions

Model	R^2	*MSE*	*RMSE*	*Training time (s)*	*Interpretability score (1–5)*
ELRM (proposed)	0.985	0.81	0.9	3.2	4.7
Random forest (RF)	0.972	1.08	1.039	18.6	2.1
Support vector regression (SVR)	0.961	1.25	1.118	24.3	1.3
LSTM	0.978	0.95	0.975	67.4	1.8
APSO-AdaBoost LSTM	0.981	0.91	0.954	112.9	1.5

Note: The interpretability score is calculated as the average of three independent scores given by education data experts, based on whether the model output supports behavioural attribution, visualises interaction effects, and provides recommendations for educational interventions.

4.6 Test of the cross-semester stability of the interaction effect coefficient in the ELRM

In response to the reliability requirements of long-term effects in educational scenarios, this study implemented a cross-semester stability verification scheme. By fitting the ELRM by semester and extracting the core interaction coefficients, the semester sequence parameter trajectory is constructed. The test process examines the interaction effects of the four LHs over a four-consecutive-teaching-cycle time span. The sliding window method is used to calculate the confidence interval and fluctuation threshold to observe how parameters drift over time. This section presents the trajectory of coefficient changes across semesters, providing a time-series basis for evaluating model generalisation. The findings are displayed in Figure 7.

Figure 7 presents the interaction effects of ELRM and TLRM on four LHs (review frequency, study frequency, class participation, and extracurricular activities) over four semesters. While the ELRM remains stable with little fluctuation in interaction effects, the TLRM demonstrates larger fluctuations, particularly in the second and fourth semesters, reflecting instability. For instance, changes in review frequency of the ELRM are steady (0.0937, 0.1225, 0.1116, and 0.1049), whereas the TLRM has much greater variation (–0.1958, 0.0248, –0.0681, and –0.2088). This illustrates that the ELRM is better able to capture the long-term interaction between LB and AP, effectively incorporating time-dependent characteristics. Conversely, the TLRM is unable to sustain stable trends, particularly in detecting intricate interactions, and results in greater fluctuations. The stability of the ELRM across semesters improves its predictive precision and generalisation, rendering it more credible in forecasting AP than the traditional model.

4.7 Expansion experiment analysis

Academic data from 12 universities across China's eastern, central, and western regions (including 'double first-class' institutions, regular undergraduate universities, and local applied universities) were collected, spanning multiple disciplines and academic years. The sample size was

expanded to 3.2 times the original dataset (N = 4,862). While maintaining the original feature system and hyperparameter configuration, we employed 10-fold cross-validation stratified by institutional type, discipline category, and academic year. We calculated R^2, MSE, RMSE, and interaction term stability coefficient (ISC) for each subgroup to evaluate the model's consistency and robustness.

Table 4 systematically evaluates the generalisation capability of ELRM across heterogeneous student groups. Data from 4,862 students spanning three types of institutions, three academic disciplines, and three grade levels demonstrate that the model maintains high prediction accuracy ($R^2 \geq 0.965$, MSE ≤ 1.08) across all subgroups, validating its cross-environmental stability. The ISC ranges between 0.82 and 0.93, reflecting consistent lagged interaction mechanisms across different groups. Science and engineering students in higher grades achieved the best performance with R^2 reaching 0.986 and MSE as low as 0.77. While freshmen and students from local applied universities showed slightly lower metrics, their performance remained within the high-precision range without significant degradation. These results strongly support ELRM's applicability in diverse educational scenarios, addressing the reviewers' concerns about the model's universality.

The study further expands systematic comparisons with representative nonlinear models, specifically including: RF, support vector regression (SVR), LSTM, and the APSO-AdaBoost LSTM that has demonstrated outstanding performance in recent educational data mining. All models were trained and tested under identical conditions (10-fold cross-validation, feature sets containing lagged variables and interaction terms, and a unified evaluation protocol) to ensure comparability. The updated results are presented in the table:

Table 5 systematically compares the predictive performance and practical characteristics of ELRM with mainstream nonlinear models under unified experimental conditions. All models were evaluated using the same feature set (including lagged terms and interaction terms) and a 10-fold cross-validation protocol. Results demonstrate that ELRM outperforms all competitors in prediction accuracy: achieving an R^2 of 0.985 and an MSE of 0.81, significantly better than RF (R^2 = 0.972), SVR (R^2 = 0.961), LSTM (R^2 = 0.978), and APSO-AdaBoost LSTM (R^2 = 0.981). Notably, ELRM requires only 3.2 seconds for training, substantially faster than LSTM-based models (over 67.4 seconds), offering significant deployment efficiency. Crucially, its interpretability score reaches 4.7 (out of 5), far surpassing black-box models (1.3–2.1), enabling educators to precisely formulate behavioural attribution and intervention strategies. These results confirm that ELRM effectively captures dynamic nonlinear characteristics in AP while maintaining the transparency of linear models through structured temporal lag and interaction mechanisms, achieving high precision, efficiency, and practicality.

5 Conclusions

This research concentrates on the dynamic analysis of factors affecting college students' AP. The study confirms that the impact of LHs and temporary learning input on AP is not isolated and static, but has significant time-varying correlations and complex synergistic mechanisms. At the level of influencing factors, the study shows that long-term stable LHs are the key foundation for the continuous improvement of AP. Temporary learning input can produce significant short-term improvement effects within a specific time window. More importantly, there is a dynamic interaction between LHs and temporary investment. High-frequency learning needs to be combined with effective review to maximise its effectiveness, while excessive extracurricular activities may compromise the effectiveness of temporary investment. In comparison with the traditional static model, the dynamic framework suggested in this study effectively captures the timeliness characteristics, time dependence, and interaction mechanism of these influencing factors. The research results have deepened the understanding of the dynamic influencing mechanisms of college students' AP, revealed the intrinsic connection between LB patterns and grade fluctuations, and provided a theoretical basis for educational administrators to design more timely and targeted learning intervention strategies.

Declarations

All authors declare that they have no conflicts of interest.

References

Al Husaini, Y. and Shukor, N.S.A. (2022) 'Factors affecting students' academic performance: a review', *Social Science Journal*, Vol. 12, No. 6, pp.284–294.

Al-Ali, R., Alhumaid, K., Khalifa, M., Salloum, S.A., Shishakly, R. and Almaiah, M.A. (2024) 'Analyzing socio-academic factors and predictive modeling of student performance using machine learning techniques', *Emerg. Sci. J.*, Vol. 8, No. 4, pp.1304–1319.

Alani, F.S. and Hawas, A.T. (2021) 'Factors affecting students academic performance: a case study of Sohar University', *Psychology and Education*, Vol. 58, No. 5, pp.4624–4635.

Amdee, N. (2024) 'A comparative study of the applicability of regression models in predicting student academic performance', *Naresuan University Engineering Journal*, Vol. 19, No. 1, pp.39–49.

Bates, S., Hastie, T. and Tibshirani, R. (2024) 'Cross-validation: what does it estimate and how well does it do it?', *Journal of the American Statistical Association*, Vol. 119, No. 546, pp.1434–1445.

Bell, C. (2025) 'Enhancing education with machine learning: predicting student readability scores', *Journal of Artificial Intelligence and System Modelling*, Vol. 3, No. 2, pp.15–31.

Crowther, P. and Briant, S. (2021) 'Predicting academic success: a longitudinal study of university design students', *International Journal of Art & Design Education*, Vol. 40, No. 1, pp.20–34.

Feraco, T., Resnati, D., Fregonese, D., Spoto, A. and Meneghetti, C. (2023) 'An integrated model of school students' academic achievement and life satisfaction. Linking soft skills, extracurricular activities, self-regulated learning, motivation, and emotions', *European Journal of Psychology of Education*, Vol. 38, No. 1, pp.109–130.

Geroldinger, A., Lusa, L., Nold, M. and Heinze, G. (2023) 'Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures – a simulation study', *Diagnostic and Prognostic Research*, Vol. 7, No. 1, p.9.

Ho, I.M.K., Cheong, K.Y. and Weldon, A. (2021) 'Predicting student satisfaction of emergency remote learning in higher education during COVID-19 using machine learning techniques', *Plos One*, Vol. 16, No. 4, p.e0249423.

Hussain, S., Gaftandzhieva, S., Maniruzzaman, M., Doneva, R. and Muhsin, Z.F. (2021) 'Regression analysis of student academic performance using deep learning', *Education and Information Technologies*, Vol. 26, No. 1, pp.783–798.

Kukkar, A., Mohana, R., Sharma, A. and Nayyar, A. (2024) 'A novel methodology using RNN+ LSTM+ ML for predicting student's academic performance', *Education and Information Technologies*, Vol. 29, No. 11, pp.14365–14401.

Kyriazos, T. and Poga, M. (2023) 'Dealing with multicollinearity in factor analysis: the problem, detections, and solutions', *Open Journal of Statistics*, Vol. 13, No. 3, pp.404–424.

Liang, G., Jiang, C., Ping, Q. and Jiang, X. (2024) 'Academic performance prediction associated with synchronous online interactive learning behaviors based on the machine learning approach', *Interactive Learning Environments*, Vol. 32, No. 6, pp.3092–3107.

Mahmud, N., Muhammat Pazil, N.S. and Azman, N.A.N. (2022) 'The significant factors affecting students' academic performance in online class: multiple linear regression approach', *Jurnal Intelek*, Vol. 17, No. 2, pp.1–11.

Mohamed, E.I., Eledum, H. and Yagoub, R. (2025) 'First-year courses and their influence on students' GPAs at the University of Tabuk using regression analysis', *Universal Journal of Educational Research*, Vol. 13, No. 1, pp.12–31.

Nieberding, M. and Heckler, A.F. (2021) 'Patterns in assignment submission times: procrastination, gender, grades, and grade components', *Physical Review Physics Education Research*, Vol. 17, No. 1, p.013106.

Pepe, M. and McCollum, J. (2023) 'The impact of using a blended learning choice model method on student performance in higher education during COVID-19', *Journal of Education for Business*, Vol. 98, No. 6, pp.324–330.

Qureshi, M.A., Khaskheli, A., Qureshi, J.A., Raza, S.A. and Yousufi, S.Q. (2023) 'Factors affecting students' learning performance through collaborative learning and engagement', *Interactive Learning Environments*, Vol. 31, No. 4, pp.2371–2391.

Salem, I.E., Alamir, A-A., Moosa, S., El-Maghraby, L., Alkathiri, N.A. and Elbaz, A.M. (2024) 'Examining different learning modes: a longitudinal study of business administration students' performance', *The International Journal of Management Education*, Vol. 22, No. 1, p.100927.

Shou, Z., Xie, M., Mo, J. and Zhang, H. (2024) 'Predicting student performance in online learning: a multidimensional time-series data analysis approach', *Applied Sciences*, Vol. 14, No. 6, p.2522.

Tagud, E.C. and Valle, A.M. (2023) 'Study habits and skills: its influence on students' academic performance', *International Journal of Research Publications*, Vol. 129, No. 1, pp.346–361.

Tao, H. (2025) 'Exploring the impact of data analysis on identifying key predictors of student performance and improving outcomes for diverse groups', *Journal of Computational Methods in Sciences and Engineering*, Vol. 25, No. 2, pp.1811–1825.

Xiong, Y., Qin, X., Wang, Q. and Ren, P. (2021) 'Parental involvement in adolescents' learning and academic achievement: cross-lagged effect and mediation of academic engagement', *Journal of Youth and Adolescence*, Vol. 50, No. 9, pp.1811–1823.

Yildiz Durak, H. (2025) 'Impact of ML-LA feedback system on learners' academic performance, engagement and behavioral patterns in online collaborative learning environments: a lag sequential analysis and Markov chain approach', *Education and Information Technologies*, Vol. 30, No. 2, pp.2623–2644.

Abbreviations

		Symbol	
AP	Academic performance	ε_t	Error term at time t
ELRM	Extended linear regression model	τ	Time period
TLRM	Traditional linear regression model	$\bar{y}$	Average of the sample
MSE	Mean squared error	$\lambda 1$	Regularisation parameter for Lasso regression
APSO-AdaBoost	Advanced Penguin search optimised adaptive boosting	$\lambda 2$	Regularisation parameter for ridge regression
LSTM	Long short-term memory	Y_t	The AP at time t
ML	Machine learning	X_{t-i}	LB variable at the i^{th} lag period
LB	Learning behaviour	Z_{t-j}	The LH variable at the j^{th} lag period
LH	Learning habit	p	Lag period.
VIF	Variance inflation factor	ϕ_i	The RC of the lag term
RC	Regression coefficient	θ_{ij}	The RC of the interaction term
RMSE	Root mean square error	X_t	The learning time
CV	Cross-validation	Z_t	The review frequency
F1	F1 score	LOOCV	Leave-one-out cross V
Z-score	Standard score	R^2	Coefficient of determination
LRM	Linear regression model	CGPA	Cumulative grade point average
RNN	Recurrent neural network	ACT	American college test