Proceedings of CONF-APMM 2025 Symposium: Controlling Robotic Manipulator Using PWM Signals with Microcontrollers
DOI: 10.54254/2753-8818/2025.AD26353

Construction and Application of a Learning Motivation Decline Prediction Model Based on Educational Big Data

Zhengyi Zhu

Department of Linguistics and Translation, City University of HongKong, HongKong, China
zhezhu6-c@my.cityu.edu.hk

Abstract. With the continuous expansion of online education, learning motivation decline has emerged as a critical issue affecting learning outcomes, while traditional assessment methods struggle to meet real-time monitoring demands. This research constructs a predictive model for learning motivation decline based on educational big data. Using the PySpark framework to process the Open University Learning Analytics Dataset (OULAD), we achieve precise prediction through a three-stage pipeline: feature engineering, model training, and system development. The feature engineering phase establishes a multi-dimensional feature system encompassing behavioral features, temporal features, and derived features. In model training, Gradient Boosted Trees (GBT) significantly outperform Random Forest (RF) with an AUC of 0.8996 versus 0.8758. The developed early-warning system supports dynamic risk-threshold adjustment and real-time intervention. Empirical analysis reveals that behavioral features contribute 85% of predictive power, with 86.8% of high-risk students having click counts <100 and 52.4% scoring below 75 points. The system response time is <1.5 seconds. The outcomes provide technical support for risk early-warning in online education and hold significant implications for optimizing pedagogical intervention strategies.

Keywords: Learning motivation decline, Gradient Boosted Trees, Educational big data, Behavioral analysis, Predictive modeling

1. Introduction

With the rapid development of Internet technology, online education has become an important development trend in the field of education by virtue of its flexibility advantage. However, in the practice of online learning, the problem of students' motivation decay is becoming more and more prominent. An in-depth analysis of the OULAD dataset shows that up to 31.16% of online learners suffer from the phenomenon of motivation decay, which leads to reduced course engagement, poor learning outcomes, and even eventual dropout. Traditional motivation assessment methods (e.g., teacher observation, student self-reporting) are subjective, inefficient, and difficult to scale up, and cannot meet the large-scale and real-time assessment needs of online education.

In the field of educational data mining and learning analytics, significant progress has been made through the integration of behavioral data with predictive models [1]. demonstrated that educational data mining techniques provide a methodological foundation for precise learning risk identification

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.AD26353

by quantifying behavioral characteristics. For instance, Pardo et al. combined self-regulated learning indicators with online engagement to construct an academic performance prediction model [2], validating the critical role of behavioral data in forecasting outcomes. Ren et al. employed multiple regression models to analyze MOOC assessment data, revealing the strong explanatory power of temporal patterns in learning behaviors for performance prediction [3]. In dropout early-warning research, Kloft et al. developed a weekly dropout prediction model using machine learning methods, achieving early risk identification through large-scale social interaction analysis [4]. Tempelaar et al. further emphasized that feedback generation effectiveness depends heavily on data feature informativeness, with dynamic behavioral indicators significantly outperforming static background data in predictive efficacy [5].

However, three critical limitations persist in current research. First, static threshold constraints remain a significant issue: most models employ fixed probability thresholds for risk determination, failing to adapt to dynamic requirements across diverse instructional contexts. Second, there is insufficient mining of dynamic behavioral features-elements such as click volatility and task completion sequences remain underexplored, leading to inadequate characterization of motivation decline processes. Third, pedagogical integration remains disconnected: despite Chen et al.'s empirical validation of the FLIPPED model's effectiveness in enhancing engagement, systematic solutions for dynamically linking pedagogical optimization with risk prediction outcomes are still lacking [6].

To address these gaps, this study proposes a dynamic risk prediction framework based on Gradient Boosted Trees (GBT). By constructing an additive model driven by behavioral dynamics and implementing a dynamic threshold rule (τ =0.6±Δ), the paper enable real-time alerts for learning motivation decline. The integrated pedagogical intervention module establishes a closed-loop system of "data collection, model prediction, strategy optimization." This approach not only inherits the predictive accuracy of traditional machine learning models but also significantly enhances adaptability and interpretability through dynamic parameter adjustment and behavioral sequence analysis.

2. Data processing and feature engineering

2.1. Data source and overview

This study employs the Open University Learning Analytics Dataset (OULAD) in Kaggle as the data source. This dataset contains rich student learning information across multiple dimensions, including basic student information, course registration data, learning behavior data, and academic performance data. After integration, a final dataset comprising 32,593 student records was formed. Student behavior data recorded 10.65 million click events, and academic assessment data contained 170,000 grade records, providing a substantial data foundation for the research.

2.2. Data cleaning and preprocessing

The raw data suffered from a variety of quality issues that significantly affected subsequent analysis and modelling. Firstly, the 'avg_score' (average score) field has a high missing rate of 18.0% (5,866/32,593 records), and the 'total_clicks' (total clicks) field has 8.75% missing values, which, if left untreated, would directly result in loss of information and compromise the accuracy of the model. Secondly, there are extreme outliers in the 'total_clicks' field, with a maximum value of 28,615, which is more than 3 standard deviations above the mean, severely deviating from the

DOI: 10.54254/2753-8818/2025.AD26353

overall data distribution and interfering with the model's ability to learn normal patterns. In addition, the clicks data show a severe right-skewed distribution (skewness = 9.2), which prevents the model from performing effective feature extraction. Meanwhile, some of the column names did not conform to the unified data processing specification (e.g., 'age, broad'), which increased the difficulty of data processing. In this study, these initial data were corrected, and the treatment strategies and effects are shown in Table 1.

Table 1. Data cleansing

Feature	Processing Strategy	Mathematical Expression	Improvement Effect
avg_score_filled	Course-average score imputation	filled=coalesce(avg_score, μ_course)	Missing rate reduced from 18.0% to 0%, ensuring academic performance data integrity.
log_total_clicks	Logarithmic transformation	ln(total_clicks + 1)	Data skewness reduced from 9.2 to 0.3, effectively improving distribution morphology.
Column Names	Unified snake_case naming	e.g., "age, broad" → "age_band"	Eliminated naming inconsistencies, reducing data processing error rate to 0%.

Key metrics comparison before and after cleaning: Overall missing rate reduced from 26.87% to 0.09%, data usability increased from 73.1% to 99.9%, outlier proportion decreased from 9.2% to 0.3%, and prediction AUC improved from 0.82 to 0.89.

2.3. Multi-dimensional feature system

The constructed multi-dimensional feature system covers behavioral, temporal, derived, academic, and demographic features. The feature processing pipeline involves: raw features entering feature engineering; separation into numerical and categorical features; numerical features undergo derived feature construction; categorical features undergo encoding; both types are vectorized; vectors are combined and normalized; finally, the processed features are input into the GBT/RF model. The formulae for the multidimensional features and the link between the features and the model are shown in Table 2.

DOI: 10.54254/2753-8818/2025.AD26353

Table 2. Data cleansing and pre-processing strategies

Feature Name	Calculation Formula	Model Association
log_total_clicks	In(total_clicks+1)	Reflects the overall learning participation of students. After logarithmic transformation, it is more in line with the normal distribution, facilitating model processing
std_weekly_clicks	$\sqrt{\sum_{i=1}^{n_i}\left(weekly_i - weekly_i^*\right)^2}$	Measures the volatility of students' weekly learning behaviors. The larger the value, the more drastic the behavior volatility
avg_click_change	$\frac{1}{n}\sum_{i=2}^{n}\frac{\text{clicks}_i - \text{clicks}_{i-1}}{\text{clicks}_{i-1}}$	Calculates the average change rate of students' weekly click volume, capturing the dynamic trend of learning participation
activity_stability	log_total_clicks / std_weekly_clicks	Comprehensively considers the total amount and volatility of learning behaviors, measuring the stability of students' learning activities. The higher the value, the more stable the behavior
learning_efficiency	avg_score_filled / total_clicks	Calculates the average score corresponding to the unit click volume, reflecting students' learning efficiency
avg_score_filled	The average score of students after filling with the course average score	Makes up for the lack of original score data, fully presenting students' academic performance
age_band_vec	One - Hot encoding of age_band	Converts students' age group information into vector form, facilitating model processing
highest_education_vec	One - Hot encoding of highest_education	Converts students' highest education information into vector form

In the feature encoding process, the raw data is first converted to a numerical index by StringIndexer, then uniquely hot encoded by OneHotEncoder, and finally assembled by VectorAssembler into a feature vector input model. This vector can be represented as:

$$\text{rawFeatures} = \begin{bmatrix} \text{log_total_clicks} \\ \text{std_weekly_clicks} \\ \vdots \\ \text{imd_band_vec} \end{bmatrix} \tag{1}$$

Next, it needs to apply Min-Max normalization to each feature in this vector.Min-Max normalization is a method of scaling the data to between 0 and 1 with the formula:

$$\text{features}_i = \frac{\text{rawFeatures}_i - \min(\text{rawFeatures})}{\max(\text{rawFeatures}) - \min(\text{rawFeatures})} \tag{2}$$

Thus, for each feature, we apply the above formula for normalization. Eventually, we obtain the normalized feature vector, which eliminates the magnitude difference between features and improves the model training efficiency and performance.

DOI: 10.54254/2753-8818/2025.AD26353

$$\text{features} = \begin{bmatrix} \log_\text{total}_\text{clicks} - \min(\text{rawFeatures}) \\ \frac{\max(\text{rawFeatures}) - \min(\text{rawFeatures})}{\text{std_weekly_clicks} - \min(\text{rawFeatures})} \\ \vdots \\ \text{imd}_\text{band}_\text{vec} - \min(\text{rawFeatures}) \\ \frac{\max(\text{rawFeatures}) - \min(\text{rawFeatures})}{\max(\text{rawFeatures}) - \min(\text{rawFeatures})} \end{bmatrix} \tag{3}$$

3. Prediction model construction

3.1. Model architecture and algorithm principles

This study selected Gradient Boosting Tree (GBT) and Random Forest (RF) as core algorithms for comparative research. The parameter configurations of the two models are shown in Table 3.

GBT: An iterative boosting algorithm based on ensemble learning. It builds multiple weak learners (decision trees) to sequentially fit the residuals of the training data, continuously reducing prediction error, ultimately forming a strong predictive model. The prediction process of GBT unfolds based on the logic of additive modelling: its final output value is not the result of a single decision tree, but a weighted summation of the predictions of all integrated trees (Equation 4). Equation (4) follows the core theoretical framework of GBT [7], which progressively approximates the optimal prediction model by iteratively fitting the residuals. This mechanism improves the prediction accuracy through gradual iterative optimisation, where each step constructs a new decision tree based on the prediction residuals of the predecessor trees, thus continuously correcting the error.

In the core parameters of the model, M =100 represents the total number of integrated trees, which are not constructed independently, but are generated iteratively in a "residual-driven" way - the training goal of each new tree is to minimise the prediction residuals of all the trees in the previous sequence, in order to gradually approach the true value. Each new tree is trained to minimise the prediction residuals of all trees in the previous sequence, in order to gradually approach the true value; γ = 0.05 is the learning rate parameter, which scales the contribution weight of a single tree to avoid excessive influence of a tree on the overall prediction, thus effectively preventing the model from losing its generalisation ability due to overfitting to the training data; and $\sum_{m=1}^{M} h_m\left(x\right)$ denotes the predicted output of the mth decision tree, whose internal node splitting follows the "gain maximisation criterion" (Eq. 5), i.e., by calculating the change in purity before and after splitting, the splitting point that can maximise the classification or prediction effect is selected.

Specifically, the algorithm calculates the "purity gain" when a node splits, the algorithm combines various factors: firstly, it evaluates the fit of the current data by the first-order gradient of the left and right subtrees (i.e., the residual sum G_L / G_R), and the smaller the residual sum is, the better it is to correct the existing error of the split. splitting approach is more effective in correcting the existing error; secondly, the stability of the data distribution is measured in combination with the second-order gradient (H_L / H_R), and the introduction of the second-order statistic allows for a more robust splitting decision and reduces the interference of extreme values; in addition, a regularisation parameter λ = 1 will be introduced, which will further inhibit the splitting gain by applying a moderate penalty to the noise in the data to avoid the model over-capturing local features and affecting the overall performance.

DOI: 10.54254/2753-8818/2025.AD26353

The original prediction F(x) generated by the model needs to be transformed by the Sigmoid function (Eq. 6) to get the final learning motor decay probability, whose function is to map the original prediction value of any real number range between 0-1 The function is to map any real range of raw prediction values to a range between 0 and 1, which is consistent with the probabilistic properties and facilitates an intuitive understanding of the level of risk.

The final risk status determination relies on the "dynamic threshold rule" (Eq. 7). Where τ is an adjustable threshold with a default value of 0.6, which can be flexibly adjusted by the teacher according to specific teaching objectives, student group characteristics or intervention strategies - for example, lowering the threshold when earlier intervention is needed for high-risk students, and raising the threshold when fewer misjudgements are desired, so as to make the risk identification more relevant to the actual teaching needs.

$$F\left(x\right)=\sum_{m=1}^{M}\gamma_m h_m\left(x\right) \tag{4}$$

$$\text{Gain}=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}\right] \tag{5}$$

$$\text{dropout_prob}=\frac{1}{1+\text{e}^{-\text{F(x)}}} \tag{6}$$

$$\text{HighRisk}=\begin{cases}1 & \text{if dropout}_\text{p}\text{rob}>0.6\text{and prediction}=1\\0 & \text{otherwise}\end{cases} \tag{7}$$

Random Forest (RF): A Bagging-based ensemble learning algorithm. It builds multiple independent decision trees through random sampling of training data and random selection of feature subsets. Final predictions are made by voting (classification) or averaging (regression). The prediction formula is (8), where $\widehat{y}$ is the final prediction, '$\hat{y}_k$' is the prediction of the 'k'-th tree, 'K' is the number of trees, and majority voting determines the final class for classification.

$$\widehat{y}=\text{mode}\{h_k\left(x\right)\}_{k=1}^{K} \tag{8}$$

Table 3. Model parameter configuration

Parameter	GBT	Random Forest
Trees (Iterations)	maxIter = 100	numTrees = 100
Max Tree Depth	maxDepth = 5	maxDepth = 5
Min Samples per Node	minInstancesPerNode = 10	minInstancesPerNode = 10
Learning Rate	stepSize = 0.05	-
Random Seed	seed = 42	seed = 42
Feature Subset Proportion	-	subsamplingRate = 0.8

DOI: 10.54254/2753-8818/2025.AD26353

3.2. Experimental design

3.2.1. Data partitioning

The dataset was partitioned using stratified sampling with the motivation_drop label as the stratification criterion, resulting in a training set (20,547 records, 63%) and a test set (12,046 records, 37%). This method's key advantage is preserving the original class distribution: the proportion of positive cases (motivation_drop= 1) was strictly maintained at 31.16% in both sets (matching the overall distribution), preventing class imbalance from random sampling. By ensuring distributional consistency, sampling bias is reduced, enhancing the model's generalization performance on unseen test data for real-world applicability. Theoretically, stratified sampling maintains class proportions, exposing models to similar data distributions during training and evaluation, thereby improving result reliability.

3.2.2. Evaluation metrics

Four core metrics were selected, calculated and analyzed during the model testing phase according to the following logic: First, a confusion matrix was generated based on the test set prediction results, then each metric value was derived. The formulas and explanations are as follows:

Area Under the ROC Curve (AUC) reflects the model's ability to distinguish between "motivation drop" (positive class) and "no drop" (negative class) across all thresholds, ranging 0-1. In this experiment, AUC=0.8996, demonstrating robustness to class imbalance (positive class: 31.16%).

Precision = $\frac{TP}{TP+FP}$ (TP: True Positives, i.e., correctly predicted high-risk students; FP: False Positives, i.e., low-risk students misclassified as high-risk). It reflects the proportion of actual high-risk students among those predicted as high-risk, reducing resource waste.

Recall = $\frac{TP}{TP+FN}$ (FN: False Negatives, i.e., high-risk students misclassified as low-risk). It reflects the proportion of actual high-risk students correctly identified, reducing missed detection risk.

F1-Score = $2 \times \frac{Precision \times Recall}{Precision+Recall}$, representing the harmonic mean of Precision and Recall to balance their trade-off and comprehensively evaluate model performance.

3.2.3. Model tuning

Hyperparameter optimization was performed using Grid Search combined with 5-fold cross-validation. The specific procedure was: First, the candidate parameter ranges (Table 3) were determined based on preliminary experiments, generating all possible parameter combinations. Subsequently, for each combination, 5-fold cross-validation was conducted on the training set (i.e., training with 4 folds and validating with 1 fold each time, repeated 5 times), recording the AUC for each validation. Finally, the average AUC for each combination was calculated, selecting the combination with the highest score as the optimal parameters.

The design principles included: limiting tree depth to maxDepth≤5 to prevent overfitting; implementing an early stopping mechanism (terminating when AUC improvement was <0.001 for 3 consecutive parameter combinations) to conserve computational resources. The effectiveness of this approach was justified by Grid Search's optimal efficiency in medium parameter spaces and 5-fold cross-validation's ability to balance bias and variance [8,9].

DOI: 10.54254/2753-8818/2025.AD26353

3.3. Experimental results

3.3.1. Model performance comparison

After rigorous tuning and testing, GBT significantly outperformed RF across all metrics.

The AUC improvement indicates GBT's superior ability to distinguish between students experiencing motivation decline and normal students. The concurrent increase in Precision and Recall shows the model reduces false positives while more comprehensively identifying true high-risk students, effectively lowering false negative risk. A comparison of the parameters of the two models is shown in Table 4.

Table 4. Model performance comparison

Evaluation Metric	GBT	Random Forest	Performance Gain
AUC	0.8996	0.8758	2.38%
Precision	0.856	0.812	5.42%
Recall	0.824	0.783	5.23%
F1-Score	0.84	0.797	5.39%

3.3.2. Feature importance analysis

Using GBT's built-in importance calculation (based on cumulative node split gain), the contribution of each feature was quantified. The formulae for the features are as follows:

$$Importance_j = \frac{1}{M} \sum_{m=1}^{M} \sum_{split\ nodes} Gain \times I(split\ uses\ feature\ j) \tag{9}$$

The top 5 behavioral features cumulatively contributed 85.7% of the predictive power, while demographic features (e.g., age, education) accounted for less than 2%, validating the dominant influence of learning behavior patterns on motivation decline. The analysis and ranking of the importance of the multidimensional features are shown in Table 5.

Table 5. Characteristic importance analysis

Feature Name	Importance Score	Importance Proportion
Log Total Clicks	0.264	28.70%
Avg Click Change	0.163	17.80%
Activity Stability	0.141	15.40%
Avg Filled Score	0.137	14.90%
Learning Efficiency	0.091	9.90%
Std Weekly Clicks	0.06	6.60%
Efficiency Index	0.053	5.70%
Age Band Vector	0.004	0.50%
Highest Education	0.004	0.50%
IMD Band Vector	0.002	0.20%

DOI: 10.54254/2753-8818/2025.AD26353

4. Learning motivation early warning system development

4.1. Layered construction

Out of 30,000 students, 5,674 were identified as medium-high risk. The system, built on a sample of 5,000 students with risk scores, follows a layered architecture (Data Layer, Processing Layer, Presentation Layer). Data Layer uses HDFS for storage; Processing Layer uses PySpark for cleaning, feature engineering, model training, and real-time prediction; Presentation Layer uses Vue.js + ECharts for an interactive UI.

Core Functionalities: Dynamic Risk Assessment: Educators customize the risk threshold 'τ' ($0.4 \leq \tau \leq 1.0$) via a slider. E.g., setting $\tau = 0.7$ instantly filters students with prediction probability > 0.7, highlighting risk levels (High/Medium/Low) in Red/Orange/Green. Provides multi-dimensional views: 3D pie charts showing risk distribution across modules; scatter plots automatically annotating high-risk points. Smart Intervention Loop: Integrates automated notifications, dynamic task adjustment based on risk level, and data export functionality for high-risk student lists and behavioral data.

4.2. Cross-dimensional analysis and intervention logic derivation

Under the concurrent access scenario of 5000 student samples, the system performance is good. The data loading time is 1.2 seconds, the response time of a single interaction is 0.8 seconds, which meets the demand for real-time decision-making; the accuracy rate of high-risk identification reaches 86.8%; and the throughput of the system is 2,000 QPS, which can support 10,000-person online education platforms.

Analysis using pie charts (min click threshold: 0, 50, 100) at $\tau = 0.6$ showed medium-high risk proportions of 28.4%, 15.9%, and 13.1% respectively, visually confirming the inverse relationship between click volume and motivation decline risk. The sharp increase in high-risk when min clicks=0 suggests non-participation or extremely low engagement.

When the threshold of the minimum number of clicks has been raised, the presentation of the histograms tends to be consistent, but from the perspective of the scatterplot, which is labelled by red/orange/green for high/medium/low risk levels, it is found that the number of people with high and potential risk, i.e., those with a recession risk value of 0.5-1, is still large. The medium and high risk values in the scatterplot are marked in orange and red, with red representing a higher than 60% risk value for motivational decay and green representing a low risk value.

DOI: 10.54254/2753-8818/2025.AD26353



Figure 1. Scatter plots (colored by risk level) (picture credit: original)

As can be seen in Figure 1, motivational decay is very much linked to learning behaviours such as clicks, with smaller clicks resulting in a higher probability of risk.

Cross-analysis of behavioral-performance clustering patterns from scatter plots and high-risk module data systematically deconstructs the core patterns of learning motivation decline and intervention logic. The student learning behaviours, such as the distribution of clicks and grades, for the three selected modules with a higher percentage of high risk are shown in Table 6. The moderate-score bracket (60-80 points) constitutes the dominant high-risk group across BBB, DDD, and FFF modules (58%-60%), aligning with the high-risk concentration in Scatter Plot Cluster 1 (clicks <100 & scores 60-80). This reveals a cognitive bias in the "moderate-score safety zone": despite reaching passing thresholds, this group experiences persistent frustration due to effort-reward mismatch stemming from fragmented learning investment (low click counts). Cluster 2 (clicks 100-500 & scores <60), as the primary high-risk cohort, exhibits extreme risk conversion rates (60%) in the FFF module, echoing the high-risk rate among low-click populations in module data and exposing deficiencies in resource accessibility and learning pathway design. In contrast, DDD demonstrates lower moderate-score risk rates (37%) and zero risk among high scorers despite a 51% high-risk rate in its low-click group, validating the efficacy of its tiered goal design and offering transferable insights for other modules.

Risk differentiation within high-click (≥500) populations reveals latent issues: Cluster 3 (clicks ≥500 & scores <60) shows minimal high-risk rates (1%-4%) across modules, indicating sufficient engagement reduces risk yet stagnant scores imply ineffective learning strategies; Cluster 4 (high-click & high-score) manifests burnout risks (e.g., FFF's 4% high-risk rate), reflecting motivation decline from excessive investment. Inter-module contradictions (e.g., BBB's low-click group accounting for 60% yet exhibiting only 39% risk rate) suggest compensatory mechanisms like forum interactions or instructor feedback, offering new perspectives for intervention design.

Through the lens of Self-Determination Theory (SDT) and Expected Value Theory (EVT), motivational decline stems from an imbalance between competence needs-expectancy assessment-learning autonomy, and this "competence needs-autonomy" imbalance is consistent with self-determination theory's core interpretation of intrinsic motivation [10], i.e., low click-throughs undermine perceptual competence; moderate scores undermine value expectations; design flaws weaken autonomy; and low click-throughs undermine value expectations. ; and design flaws weaken

DOI: 10.54254/2753-8818/2025.AD26353

autonomy. Interventions must stratify by cluster characteristics: implement progress monitoring and task decomposition for Cluster 1 ("crammers") to counteract fragmented engagement; deploy diagnostic assessments and adaptive pathways for Cluster 2 ("foundationally weak learners") to address competence gaps; provide metacognitive training for Cluster 3 ("inefficient investors") to rebuild learning strategies; optimize workload management with enrichment tasks for Cluster 4 ("high-achieving burnout"). Concurrently, establish dual-threshold alerts (clicks ≥500 & scores ≥80), while lowering click-alert thresholds for high-risk modules like FFF (e.g., to 300 clicks) to construct localized parameter systems. This dismantles cognitive blind spots around moderate-score risks and strengthens formative assessment's positive regulation of motivation.

In essence, motivation decline manifests as a triadic imbalance among "expectation, competence, and behavior." Learners with moderate scores constitute the largest risk group, primarily due to cognitive biases, such as misperceiving passing scores as a "safe zone" and low-engagement behaviors (e.g., fewer than 500 clicks). Meanwhile, significant disparities exist across modules (for instance, the FFF module shows a 60% risk rate), highlighting the need for localized alert parameters. By addressing technical accessibility issues, optimizing tiered goals, and reinforcing real-time feedback, dynamic maintenance of learning motivation can be effectively achieved.

Table 6. High-risk module analysis

Module	Analysis Category	Students	High-Risk Students	Category %	High-Risk %
BBB	Clicks <500	765	298	60%	39%
BBB	Clicks ≥500	520	4	41%	1%
BBB	Score <60	105	4	8%	4%
BBB	Score 60-80	742	290	58%	39%
BBB	Score >80	438	8	34%	2%
DDD	Clicks <500	411	208	42%	51%
DDD	Clicks ≥500	558	11	58%	2%
DDD	Score <60	227	9	23%	4%
DDD	Score 60-80	565	210	58%	37%
DDD	Score >80	177	0	18%	0%
FFF	Clicks <500	308	185	25%	60%
FFF	Clicks ≥500	924	36	75%	4%
FFF	Score <60	104	10	8%	10%
FFF	Score 60-80	741	196	60%	27%
FFF	Score >80	387	15	31%	4%

5. Application value

The main findings of the study show that the dynamic features of learning behaviours are absolutely dominant in predicting the decline of learning motivation, and their overall feature importance is as high as 85.7%, which strongly supports the theoretical view of "behaviour-driven motivation". Meanwhile, the study also identifies a rule with high practical value - "less than 500 clicks and an average score of less than 75", which provides a clear, quantifiable and rapid risk screening criterion for online education platforms.

DOI: 10.54254/2753-8818/2025.AD26353

At the practical application level, the system developed in this study brings significant value to different educational participants: teachers can use the system to pinpoint high-risk students in just 10 minutes, and design more targeted intervention strategies by using the visual behavioural data provided by the system to gain insights into the specific problems, empirical studies have confirmed that targeted intervention strategies designed in conjunction with behavioural heat mapping can enhance sustained learning rates [11], significantly outperforming traditional methods, thus dramatically improving the efficiency of interventions; educational administrators should attach great importance to the results of the study in terms of the importance of the characteristics revealed by the study. Educational administrators should prioritize the construction and optimization of learning behavior monitoring and analysis tools, as they are the key infrastructure for identifying and responding to motivation problems; and students can receive real-time feedback on their learning status from the system, which not only helps them to self-regulate, but also allows teachers to intervene in a timely manner to prevent further decline of motivation and safeguard the effectiveness of learning.

6. Conclusions

This study successfully constructs a high-precision prediction model of learning motivation decline with the help of educational big data and gradient boosting tree algorithm, with an AUC value of 0.8996, an achievement that strongly confirms the key influence of learning behavioural characteristics. Meanwhile, the developed early warning system is capable of precise risk identification with an accuracy rate of 86.8%, and also has the ability to quickly respond to interventions with a response time of less than 1.5 seconds, and can provide data-driven decision support, which provides a practical solution for improving the quality of online education.

Going forward, this research will further deepen applications around the key challenges that exist today. In terms of real-time computational capability, although existing systems are already capable of rapid response, in the face of massive and continuously growing educational data streams, the future needs to break through the bottleneck of more efficient and real-time streaming computation technology based on the discrete streaming computation paradigm, replacing the existing batch processing model to eliminate the latency constraints, and ultimately realising the millisecond-level analysis and early warning of the dynamic learning state. The final goal is to achieve millisecond-level analysis and early warning of dynamic learning states. Regarding model interpretability, although the gradient boosting tree model has high prediction accuracy, its "black box" feature limits educators' understanding and trust in the inner logic of the prediction results. Therefore, Future research will focus on the in-depth integration of interpretable techniques: adopting the SHAP unified interpretation framework to quantify key features such as weekly click-through rate fluctuations of >35% to pry open the black box of the model; supplemented with LIME to intuitively reveal the core drivers and transmission paths of motivational decay, which will not only enhance the credibility of the early warning results, but also provide a targeting hand for precise interventions. In addition, in terms of quantitative assessment of intervention effects, the current system focuses on early identification of risks and intervention triggers, but lacks a systematic quantitative assessment framework for the actual effects of interventions. In the future, we will design a rigorous empirical research programme, establish scientific assessment indexes and methods, quantitatively analyse the effectiveness and long-term impacts of different intervention strategies, and form a closed-loop feedback mechanism of "Early Warning-Intervention-Assessment-Optimisation". In the future, we will design rigorous empirical research programmes, establish scientific evaluation indicators and methods to quantitatively analyse the effectiveness and long-

DOI: 10.54254/2753-8818/2025.AD26353

term impact of different intervention strategies, and form a "warning-intervention-assessment-optimisation" closed-loop feedback mechanism to continuously improve the accuracy and efficiency of intervention.

By solving the three core challenges of real-time calculation, model transparency and quantification of intervention effects, this study will promote the application of big data in education to develop in a more intelligent, accurate and practical guidance direction.

References

[1] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer.
[2] Pardo, A., Han, F., & Ellis, R. A. (2017). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. IEEE Transactions on Learning Technologies, 10(1), 82-92.
[3] Ren, Z., Rangwala, H., & Johri, A. (2016). Predicting performance on MOOC assessments using multi-regression models. arXiv preprint arXiv: 1605.02269.
[4] Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. Proceedings of the EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs, 60-65.
[5] Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation. Computers in Human Behavior, 47, 157-167.
[6] Chen, Y., Wang, Y., & Chen, N. S. (2014). Is FLIP enough? Or should we use the FLIPPED model instead? Computers & Education, 79, 16-27.
[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
[8] Zaharia, M., et al. (2013). Discretized Streams... SOSP'13.
[9] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach... NIPS.
[10] Ryan, R. M., & Deci, E. L. (2000). Self-determination theory... American Psychologist.
[11] Li, J. C., et al. (2020). Empirical study on online learning behavior analysis and intervention strategies Distance Education in China.