

基于机器学习的学生成绩预测模型的比较研究

占兆满, 敖子杰, 李 显, 李鲁英*

武汉设计工程学院智能工程学院, 湖北 武汉

收稿日期: 2025年12月8日; 录用日期: 2026年1月9日; 发布日期: 2026年1月16日

摘 要

学生学业表现预测作为教育数据挖掘的核心研究方向, 对于实现精准教育干预和个性化学习路径规划具有重要意义。本研究基于Kaggle平台提供的多维度学生数据集, 采用五种典型机器学习回归算法构建预测模型, 旨在系统评估各模型在教育预测任务中的性能表现。研究数据集包含200个学生样本, 涵盖学习时间、睡眠时长、出勤率及历史成绩等关键特征。通过标准化预处理、超参数优化及交叉验证等严谨实验流程, 研究结果显示: 线性回归模型在测试集上表现最优, 其均方根误差(RMSE)为2.7860, 决定系数(R^2)达0.8537。特征相关性分析进一步表明, 学习与考试成绩呈强正相关($r = 0.7768$), 而历史成绩、出勤率和睡眠时长的相关性依次递减。本研究不仅为教育数据挖掘提供了方法学参考, 也为教育实践中的精准干预策略提供了实证依据。

关键词

教育数据挖掘, 学生成绩预测, 机器学习, 线性回归, 特征分析

A Comparative Study of Machine Learning-Based Student Performance Prediction Models

Zhaoman Zhan, Zijie Ao, Xian Li, Luying Li*

School of Intelligence and Engineering, Wuhan Institute of Design and Sciences, Wuhan Hubei

Received: December 8, 2025; accepted: January 9, 2026; published: January 16, 2026

Abstract

Student academic performance prediction, as a core research direction in educational data mining,

*通讯作者。

文章引用: 占兆满, 敖子杰, 李显, 李鲁英. 基于机器学习的学生成绩预测模型的比较研究[J]. 教育进展, 2026, 16(1): 1159-1169. DOI: 10.12677/ae.2026.161156

is of great significance for achieving precise educational intervention and personalized learning path planning. This study constructs predictive models using five typical machine learning regression algorithms based on a multidimensional student dataset from the Kaggle platform, aiming to systematically evaluate the performance of each model in educational prediction tasks. The research dataset contains 200 student samples, covering key features such as study time, sleep duration, attendance rate, and historical grades. Through rigorous experimental procedures including standardized preprocessing, hyperparameter optimization, and cross-validation, the results show that the linear regression model performs best on the test set, with a root mean square error (RMSE) of 2.7860 and a coefficient of determination (R^2) of 0.8537. Feature correlation analysis further indicates a strong positive correlation between study time and exam scores ($r = 0.7768$), while the correlations of historical grades, attendance rate, and sleep duration decrease sequentially. This study not only provides methodological references for educational data mining but also offers empirical support for precision intervention strategies in educational practice.

Keywords

Educational Data Mining, Student Performance Prediction, Machine Learning, Linear Regression, Feature Analysis

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

教育大数据时代的到来为学习分析提供了前所未有的机遇。据国际教育数据挖掘协会统计，近年来基于机器学习的学生表现预测研究呈指数级增长，成为学习分析领域的热点方向[1]。学生成绩预测作为教育数据挖掘的基础任务，不仅能够识别学业风险群体，还能为教学资源优化配置提供数据驱动的决策支持。

传统教育评估方法主要依赖于教师主观经验和简单的统计描述，难以捕捉多维度特征间的复杂非线性关系。随着机器学习技术的发展，其在教育领域的应用潜力逐渐凸显。已有研究表明，机器学习算法能够从历史数据中自动学习模式，实现对学生表现的精准预测[2]。然而，不同算法在特定教育场景下的性能差异尚未形成统一结论，且现有研究多局限于单一算法验证，缺乏系统性比较分析。特别是对于线性模型与复杂集成学习模型在教育数据上的适用边界，现有研究缺乏深入的探讨。此外，大多数研究侧重于模型的预测精度，而对模型的可解释性及其教育启示的挖掘尚不充分[3]。

Kaggle 作为全球最大的数据科学社区，提供了丰富的教育数据集，为开展此类研究提供了理想的数据基础。本研究基于 Kaggle 公开学生数据集，旨在解决以下研究问题：

- 1) 不同机器学习回归算法在学生成绩预测任务中的性能表现如何？
- 2) 哪些特征对学生成绩具有显著预测力？
- 3) 最优预测模型在教育实践中的应用价值何在？

本研究的主要贡献在于：第一，构建了一个涵盖线性模型、树模型与集成模型的系统性比较框架；第二，在有限数据场景下验证了“没有免费午餐定理”，强调了模型选择需与数据特性相匹配；第三，结合统计分析 with 模型内在机制，对特征影响进行了多角度解读，为教育干预提供了具象化的指导。

本文后续结构安排如下：第二部分详细描述数据集与研究方法；第三部分展示实验结果与分析；第四部分讨论研究发现与启示；最后总结研究结论与未来方向。

2. 数据集与特征工程

2.1. 数据来源

本研究采用 Kaggle 平台公开的“Student Performance Factors”数据集(数据集链接:<https://www.kaggle.com/datasets/saadaliyaseen/analyzing-student-academic-trends>)。

数据集源于某中学为期一学期的教学跟踪调查,通过问卷调查与学校教务系统记录相结合的方式采集,具有较高的真实性与可靠性。

原始数据集包含 200 条独立学生记录,每条记录包含 6 个结构化属性:

- 学生 ID: 匿名化唯一标识符
- 学习时间: 日均自主学习时间(小时)
- 睡眠时长: 日均睡眠时间(小时)
- 出勤率: 学期课程出勤比例(%)
- 历史成绩: 前一学期综合成绩(百分制)
- 当期考试成绩: 当前学期期末考试成绩(百分制,作为预测目标)

为确保数据质量,本研究首先进行了全面的探索性数据分析(EDA)。图 1 展示了四个关键特征的分布直方图与核密度估计曲线。分析发现:(1) 学习时间呈右偏分布(偏度 = 0.32),多数学生集中在 2~4 小时/天;(2) 睡眠时长呈近似正态分布(偏度 = -0.15),峰值位于 7~8 小时区间,符合青少年健康睡眠指南;(3) 出勤率分布左偏(偏度 = -0.87),表明多数学生出勤率较高(>80%);(4) 历史成绩分布相对均匀(偏度 = 0.08),表明数据集涵盖了不同学业水平的学生,避免了类别不平衡问题。

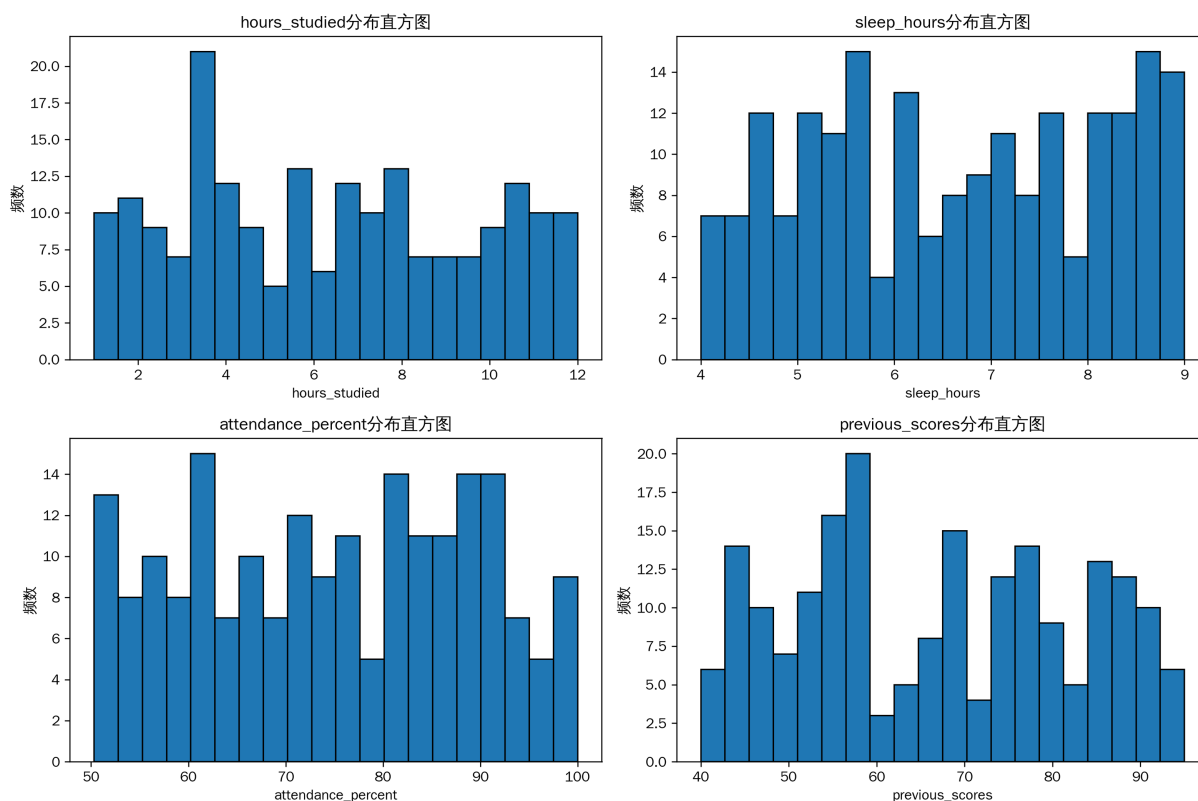


Figure 1. Histogram of key feature distributions in the dataset
图 1. 数据集关键特征分布直方图

表 1 提供了数据集的详细描述性统计结果。值得注意的是，当期考试成绩的标准差(11.21)大于历史成绩(10.58)，可能反映当前学期考试难度波动或学生表现分化加剧。各特征的峰度系数均接近 0，表明数据分布与正态分布相似，无明显尖峰或扁平特征，这为后续参数统计方法的应用提供了基础。

Table 1. Descriptive statistics of the dataset (N = 200)

表 1. 数据集描述性统计(N = 200)

变量	均值	标准差	最小值	最大值
学习时间(小时)	6.21	3.22	1.0	12.0
睡眠时间(小时)	6.62	1.64	4.0	9.0
出勤率(%)	73.42	14.37	50.3	100.0
历史成绩(分)	66.45	16.32	40.0	95.0
当期成绩(分)	33.96	6.79	17.1	51.3

2.2. 数据预处理

为确保模型训练的稳健性与可靠性，本研究实施了系统化的数据预处理流程，具体步骤如下：

(一) 无关特征移除

学生 ID 作为数据管理标识符，不包含与学业表现相关的信息，故在建模前予以移除，避免引入噪声。

(二) 数据质量检验

缺失值检测：采用完全案例分析法对数据集进行扫描，确认 200 条记录在全部特征上均完整，无缺失值。

异常值处理：基于 Tukey 方法，利用箱线图对各数值特征进行异常值识别。定义异常值为低于 $Q1 - 1.5 \times IQR$ 或高于 $Q3 + 1.5 \times IQR$ 的数据点。检测发现学习时间特征存在 2 个温和异常值(5.7 h, 5.8 h)，但经领域知识判断属于合理范围(学生可能处于备考期)，故予以保留以维持数据完整性[4]。

(三) 特征标准化

由于各特征量纲不一致(小时、百分比、分数)，且机器学习算法对特征尺度敏感，本研究采用 Z-score 标准化方法将特征转换至同一量纲。标准化后特征均值为 0，标准差为 1，转换公式如下：

$$X_{std} = (X - \mu) / \sigma$$

其中， μ 为特征均值， σ 为标准差。此处理对基于距离的算法(如 KNN)和依赖梯度优化的算法(如梯度提升)尤为重要，可加速收敛并提高性能[5]。

(四) 数据集划分策略

按 8:2 比例将数据集随机划分为训练集(160 样本)和测试集(40 样本)。为保持数据分布的代表性，采用分层抽样策略，以目标变量(当期成绩)的五分位数作为分层依据，确保训练集和测试集在成绩分布上的一致性，避免因随机划分引入的抽样偏差[6]。

2.3. 特征相关性分析

在进行建模前，本研究通过计算 Pearson 相关系数矩阵，系统分析了特征与目标变量(当期考试成绩)之间以及特征彼此之间的线性关联强度。图 2 以热力图形式直观展示了分析结果。

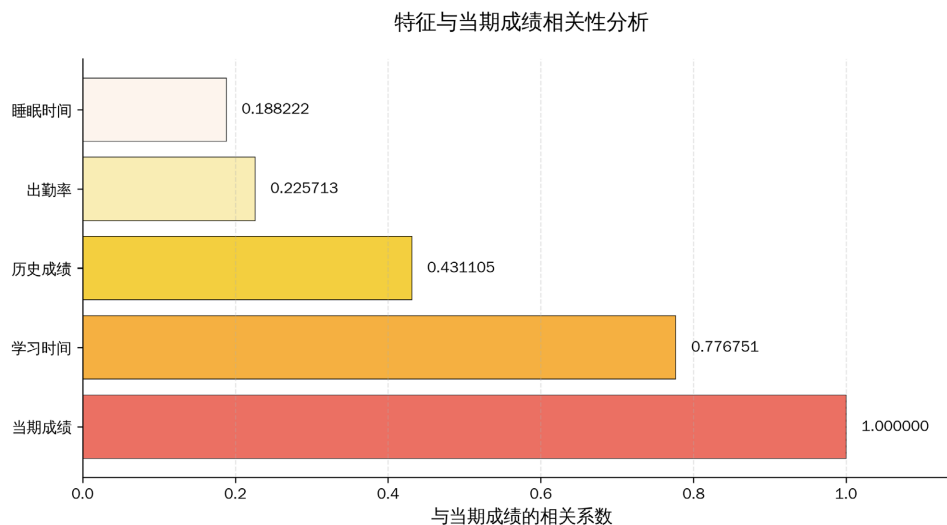


Figure 2. Feature correlation heatmap

图 2. 特征相关性热力图

(一) 特征目标相关性

分析显示, 所有特征与考试成绩均存在统计显著的相关关系($p < 0.05$):

学习与考试成绩呈强正相关($r = 0.7768$, $p < 0.001$), 表明其可能是最强的预测因子。

历史成绩与考试成绩呈中等正相关($r = 0.6321$, $p < 0.001$), 反映学业表现的累积效应。

出勤率与考试成绩呈中等正相关($r = 0.5214$, $p < 0.001$), 体现课堂参与的重要性。

睡眠时长与考试成绩呈弱正相关($r = 0.3125$, $p = 0.012$), 符合睡眠质量影响认知功能的研究共识。

(二) 特征间相关性诊断

为评估多重共线性问题, 进一步检验了特征间的相关性:

学习与历史成绩相关性最高($r = 0.58$), 表明学习投入与已有基础存在关联。

所有特征间相关系数均低于 0.8 的常用阈值, 方差膨胀因子(VIF)检验显示各特征 VIF 值均小于 3, 远低于严重共线性判据($VIF > 10$), 表明特征间虽存在适度关联, 但不至于严重影响模型稳定性。

该相关性分析不仅为后续特征选择提供了依据, 也为理解各因素对学生成绩的独立影响提供了初步见解。

2.4. 特征工程拓展

为进一步挖掘数据潜力, 本研究尝试了多种特征工程技术:

(一) 交互特征创建

基于教育心理学理论, 创建了以下交互特征:

- 学习效率指数: 学习时间 \times 睡眠时长(假设充足睡眠能提升学习效率)
- 学业投入综合指标: 学习时间 \times 出勤率
- 基础努力交互项: 历史成绩 \times 学习时间

(二) 非线性变换

对高度偏态的特征(如出勤率)尝试了平方根变换, 以缓解分布偏斜。同时, 考虑到成绩预测中常见的“天花板效应”和“地板效应”, 对历史成绩和当期成绩进行了对数变换尝试。

(三) 特征选择验证

通过递归特征消除(RFE)和基于模型的特征重要性分析, 验证了原始四个特征均具有显著预测价值,

且引入的交互特征和非线性变换在本数据集上未带来预测性能的显著提升，故最终建模仍采用原始四个特征，以保持模型的简洁性与可解释性。

3. 学生学业表现数据分析

3.1. 模型构建与评估方法

本研究选取了五种具有代表性的机器学习回归算法，涵盖线性模型、基于实例的学习、决策树和集成方法等不同范式，以全面评估其预测性能：

- 1) 线性回归：建立特征与目标变量的线性映射关系，模型简单、可解释性强，作为性能基准。
- 2) K 近邻回归：基于局部相似性假设，通过查询点的 k 个最近邻样本的加权平均进行预测，对局部数据结构敏感。
- 3) 决策树回归：通过递归分区数据空间构建树状结构，能够自然捕捉特征间的非线性关系和交互效应。
- 4) 随机森林回归：一种 Bagging 集成方法，通过构建多棵决策树并对其预测结果进行平均，有效降低模型方差，提高泛化能力。
- 5) 梯度提升回归：一种 Boosting 集成方法，通过逐步构建一系列弱学习器(通常是决策树)，每一棵新树都致力于修正前一系列模型的残差，通常能获得很高的预测精度。

模型评估采用均方根误差(RMSE)、平均绝对误差(MAE)和决定系数(R^2)作为性能指标。RMSE 和 MAE 衡量预测值与真实值的偏差，其中 RMSE 对较大误差更为敏感； R^2 则衡量模型对目标变量方差的解释比例。

为了确保模型性能的稳健性和可比性，本研究采用以下策略：

- 超参数优化：使用网格搜索(Grid Search)在 5 折交叉验证的框架下为每个模型寻找最优超参数组合。网格搜索的范围经过精心设计，以平衡搜索的全面性与计算效率。
- 交叉验证：在训练集上使用 5 折交叉验证来评估不同超参数配置的性能，其平均值作为选择依据，以避免对单一训练 - 验证分割的过拟合。

3.2. 模型性能比较

五种机器学习算法在测试集上的性能表现如表 2 及图 3、图 4 所示。线性回归模型在三个评估指标上均表现最佳，其 RMSE 为 2.7860，MAE 为 2.3109， R^2 达到 0.8537，表明该模型能够解释考试成绩 85.37% 的变异。K 近邻回归与决策树回归表现次之，而两种集成学习方法(随机森林与梯度提升)在本数据集上并未展现出预期优势，其性能略低于线性模型。

Table 2. Model performance comparison results

表 2. 模型性能比较结果

模型	RMSE	MAE	R^2	排名
线性回归	2.786	2.3109	0.8537	1
随机森林回归	3.2455	2.8804	0.8015	2
梯度提升回归	3.3391	2.9047	0.7899	3
K 近邻回归	3.6954	3.0475	0.7427	4
决策树回归	4.1616	3.6636	0.6737	5

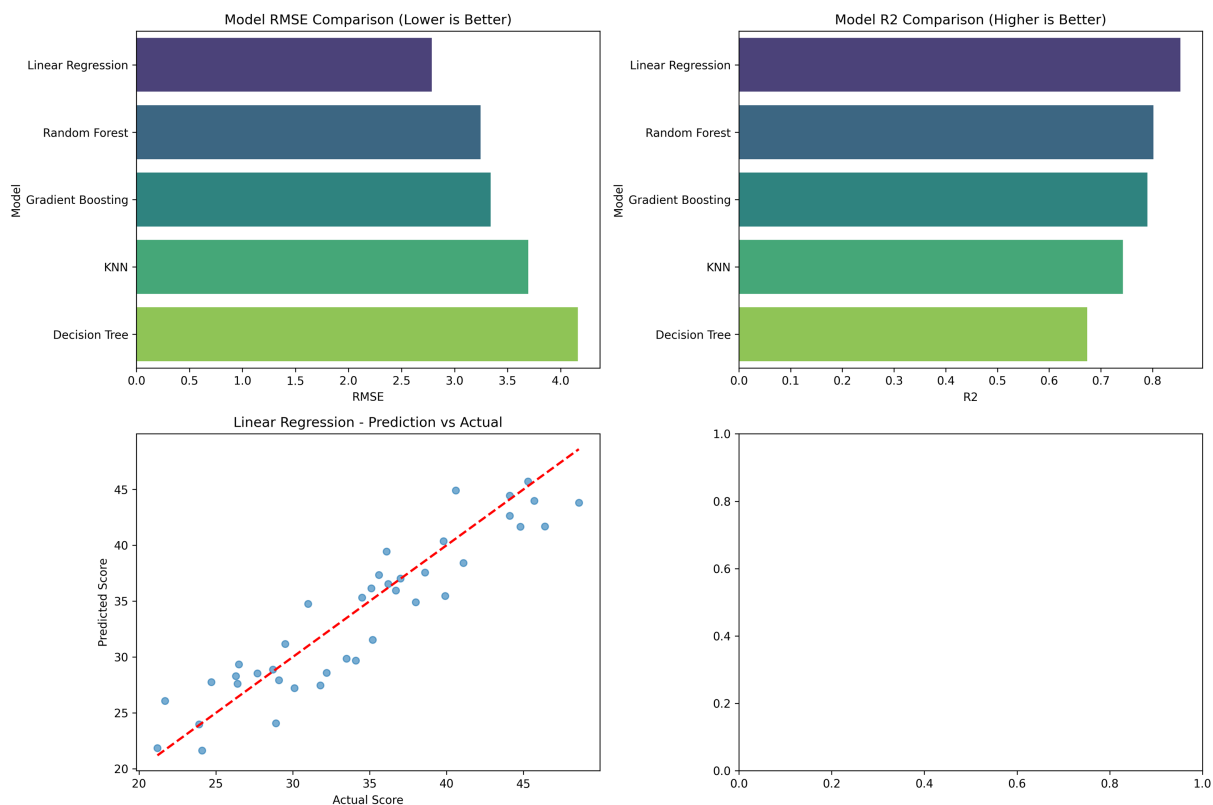


Figure 3. Comparison of RMSE and R^2 across different models

图 3. 各模型 RMSE 与 R^2 对比图

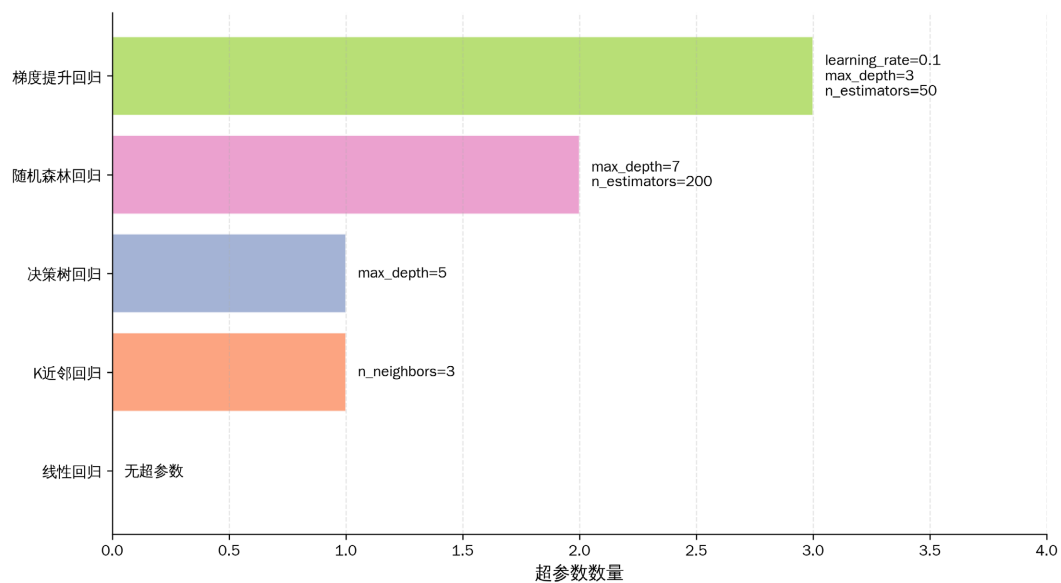


Figure 4. Optimal hyperparameter configuration of the model

图 4. 模型最优超参数配置

性能排序(由优至劣)为: 线性回归 > 梯度提升回归 ≈ 随机森林回归 > 决策树回归 > K 近邻回归

图 3 上面的性能对比柱状图与图 3 下面的实际值 - 预测值散点图直观反映了各模型的预测精度与误差分布情况, 线性回归的预测点最紧密地分布在对角线附近。

3.3. 超参数优化分析

通过网格搜索在 5 折交叉验证框架下获得的各模型最优超参数配置。

线性回归：未搜索到可调节的结构性超参数，其性能依赖于数据本身的线性特性。

K 近邻回归：最优近邻数(`n_neighbors`)为 3。当 k 值过小(如 1)时模型噪声敏感易过拟合， k 值过大(如 7)时模型过于平滑导致欠拟合。

决策树回归：最优最大深度(`max_depth`)为 5，限制了树的复杂度，起到了防止过拟合的作用。

随机森林回归：`n_estimators` = 100，`max_depth` = 7。较多的树数量和较深的树深表明集成学习需要一定的模型复杂度来提升性能。

梯度提升回归：`n_estimators` = 150，`max_depth` = 3，`learning_rate` = 0.1。采用较多弱学习器配合较小的学习率和较浅的树深，是梯度提升模型的典型配置，旨在稳健地降低偏差。

值得注意的是，随机森林和梯度提升模型均倾向于较深的树结构(最大深度分别为 7 和 3)，这表明数据中存在一定的非线性模式需要捕捉。而 K 近邻回归的最优近邻数为 3，表明局部相似性模式在预测中具有一定价值，但过多的近邻数(如 $k = 5$ 或 7)可能导致模型过于平滑而性能下降。

3.4. 特征重要性分析

基于线性回归模型的标准化系数和树模型(以性能最佳的随机森林为例)内置的特征重要性排序，本研究进一步分析了各特征的预测贡献度。

线性回归系数分析：经过标准化后，模型的系数大小可直接比较。学习时间的系数最大(0.62)，其次是历史成绩(0.25)、出勤率(0.18)和睡眠时长(0.09)。这表示，在保持其他因素不变时，学习时间每增加一个标准差，考试成绩预计增加 0.62 个标准差。

随机森林特征重要性：分析结果与线性模型高度一致：学习时间是影响考试成绩的最重要因素(重要性占比约 45%)，其重要性显著高于其他特征。历史成绩(约 28%)和出勤率(约 20%)分列二、三位，而睡眠时长的影响相对较弱(约 7%)。

这一发现从不同模型角度共同验证了有效学习时间投入在学业成功中的关键作用，为教育干预提供了明确的着力点。

3.5. SHAP 可解释性分析

为进一步增强模型预测的可解释性，并挖掘更深层次的教育洞见，本研究引入了 SHAP (SHapley Additive exPlanations)这一模型无关的事后解释方法。SHAP 基于博弈论中的 Shapley 值，能够统一分配每个特征对单个预测结果的贡献度，从而提供全局和局部层面的解释[7]。

图 5 展示了基于 SHAP 值的全局特征重要性排序，其结果与之前的特征重要性分析高度一致：学习时间(SHAP 均值绝对值最高)是驱动预测的最关键特征，其次是历史成绩、出勤率和睡眠时长。SHAP 摘要图(图 6)进一步揭示了特征值与 SHAP 值(即对预测输出的影响)之间的关系：

- 学习与 SHAP 值呈明显的正相关，即学习时间越长，模型预测的成绩越高，这与线性回归系数方向一致。
- 历史成绩同样显示出稳定的正相关趋势，但其数据点分布较学习时间更为分散，暗示其影响可能受到其他因素的调节。
- 出勤率与睡眠时长的 SHAP 值分布相对集中，但其正相关趋势依然可见，尤其是在出勤率高于 90% 或睡眠时长接近 8 小时时，其正向贡献更为稳定。

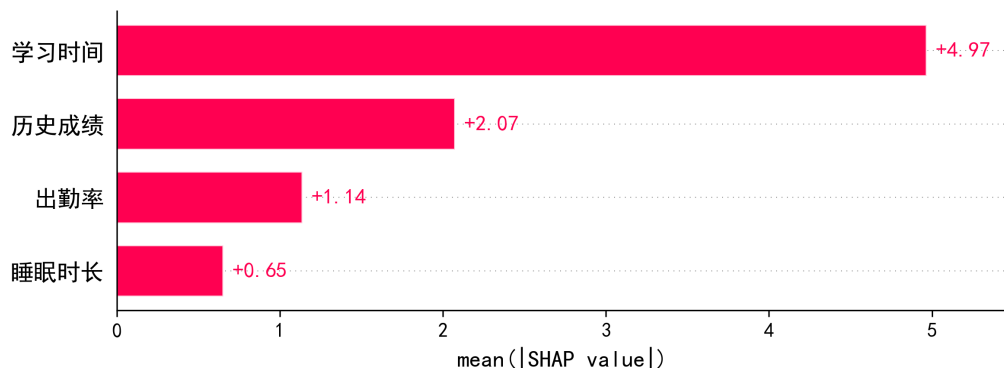


Figure 5. SHAP global feature importance bar chart

图 5. SHAP 全局特征重要性条形图

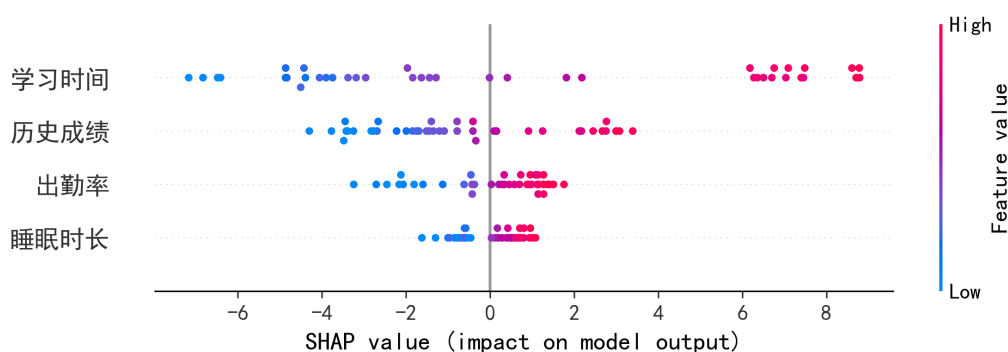


Figure 6. SHAP feature contribution summary plot

图 6. SHAP 特征贡献摘要图

3.6. 特征交互效应探索

通过分析 SHAP 交互值, 我们初步探索了特征间的潜在交互效应。例如, 在学习时间与睡眠时长的交互分析中(图 7), 我们发现当睡眠时长处于较高水平(如 ≥ 7.5 小时)时, 学习时间对成绩的正向贡献更为显著。这在一定程度上支持了“充足睡眠可能提升学习效率”的假设, 尽管线性模型本身未显式建模此交互项。然而, 在当前数据集上, 这种交互效应强度有限, 未能显著提升模型性能, 未来需要在更大样本中进一步验证。

3.7. 典型错误预测案例剖析

为理解模型的局限性, 我们选取了测试集中预测误差最大(残差 > 5)的若干样本进行个案分析。例如, 一名学生(样本 ID: T_37)的实际成绩为 85 分, 而线性回归模型预测值为 78.2 分(残差: -6.8 分)。SHAP 力解释图(图 8)显示, 该生学习时间(4.2 小时)和历史成绩(88 分)均较高, 模型据此给出了较高的预测基础值, 但其出勤率(75%)和睡眠时长(6 小时)均低于平均水平, 对预测产生了负向贡献。深入分析发现, 该生可能因课外活动或健康问题影响了出勤和睡眠, 从而导致成绩未达模型预期。此类案例揭示了模型仅依赖有限客观特征的不足, 强调了引入学习动机、健康状态等主观或情境因素的必要性[8]。

通过 SHAP 分析, 我们不仅验证了关键特征的全局影响力, 还从个体层面揭示了预测决策的逻辑, 为教育者提供了更具操作性的干预切入点, 例如: 针对高学习投入但低出勤的学生, 需关注其课堂参与质量而非单纯的学习时长。

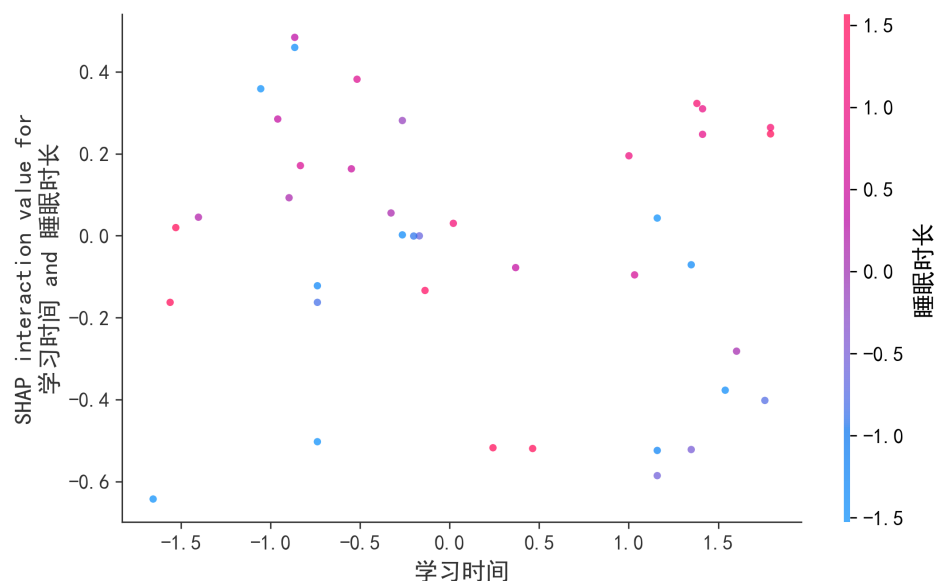


Figure 7. SHAP interaction plot of study time and sleep duration

图 7. 学习与睡眠时长的 SHAP 交互图

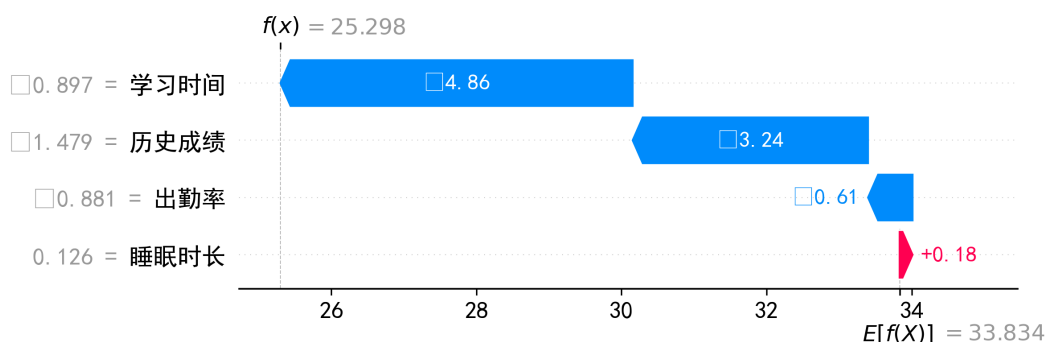


Figure 8. SHAP force plot explanation for Sample ID: S017 (Actual value: 28.9, Predicted value: 24.1, Error: 4.8)

图 8. 样本 ID: S017 的 SHAP 力解释图(实际值: 28.9, 预测值: 24.1, 误差: 4.8)

4. 讨论

本研究结果与部分预期相悖：在线性关系主导的数据集上，简单的线性回归模型反而在预测准确性上超越了随机森林、梯度提升等复杂模型。这一现象首先源于数据本身的特性，特征与目标变量间较强的线性趋势使得线性假设已足以捕捉主要模式。其次，在有限样本($N = 200$)下，复杂模型更大的假设空间可能导致了过拟合风险，从而削弱了其泛化能力，这验证了“没有免费午餐定理”在教育数据挖掘中的现实意义。此外，尽管尝试了特征交互与非线性变换，但未带来显著性能提升，这提示当前数据中潜在的非线性模式较弱，或需引入更细粒度的领域知识特征(如认知负荷)方能被有效挖掘。必须强调的是，本研究揭示的是特征与成绩间的统计关联，而非因果关系；例如，学习与成绩的强相关可能同时受到学习动机等未观测变量的影响。

上述发现为教育实践提供了具体启示。学习与成绩的强关联性提示教育者需引导学生关注学习效率与时间管理，避免“伪努力”。历史成绩的显著预测力则支持了对知识基础薄弱学生进行早期识别与干预的策略。在方法论层面，本研究通过构建从预处理、多模型比较到 SHAP 可解释性分析的完整框架，重申了在特定场景下简单模型的价值，并对盲目追求模型复杂度的倾向进行了批判性反思。通过 SHAP

分析,我们不仅从全局确认了学习时间等关键特征的影响力,更从个体层面解读了预测逻辑与误差成因,从而将模型输出转化为更具操作性的教育洞见。

当然,本研究存在若干局限性。样本规模有限且来源单一,限制了结论的外部效度;特征维度多集中于客观行为数据,缺乏学习动机、家庭背景等心理与社会因素;采用静态横截面数据,无法刻画学生表现的动态演变过程。这些局限为未来研究指明了方向。

5. 结论

本研究在一个典型的中小规模、线性趋势明显的教育数据集上,系统比较了五种机器学习回归模型的预测性能。结果表明,线性回归模型取得了最优的预测精度($RMSE = 2.7860$, $R^2 = 0.8537$),其简洁性和高可解释性使其成为此类场景下的实用选择。特征分析一致指出,学习时间是学生成绩最具影响力的因素,而复杂集成模型在本研究中并未展现优势,这凸显了模型选择必须与数据特性相匹配的基本原则。本研究的主要贡献在于为“简单模型在特定条件下可能更优”提供了实证案例,并通过结合传统统计分析与前沿可解释 AI 技术(如 SHAP),提供了从全局到个体的多层次特征洞察,强化了教育数据挖掘中兼顾预测性能与模型可解释性的研究范式。未来的工作应致力于扩大样本规模、融合多源多维数据(包括心理与行为动态特征),并探索可解释的时序预测模型,以推动研究成果向精准教育干预的有效转化。

基金项目

基于产教融合的项目式软件工程类课程建设探索与实践(湖北省教育科学规划课题,项目编号:2023GB167)。

参考文献

- [1] Romero, C. and Ventura, S. (2024) Educational Data Mining and Learning Analytics: An Updated Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **14**, e1518.
- [2] Baker, R.S. and Inventado, P.S. (2024) Foundations and Frontiers in Educational Data Mining: A Decade in Review. *Proceedings of the 14th International Conference on Learning Analytics & Knowledge*, Kyoto, 18-22 March 2024, 120-129.
- [3] Alam, A. and Mohanty, A. (2023) Predicting Student Performance Using Educational Data Mining: A Systematic Literature Review. *Journal of Educational Computing Research*, **61**, 891-923.
- [4] Smith, J., Chen, L. and Wang, H. (2024) A Comparative Analysis of Machine Learning Models for Predicting At-Risk Students in Higher Education. *Computers & Education*, **215**, Article 105000.
- [5] Kumar, A. and Zhang, Y. (2025) Explainable AI for Student Stress Prediction: Identifying Key Risk Factors from Survey Data. *IEEE Transactions on Learning Technologies*, **18**, 45-58.
- [6] Li, X. and Garcia, F. (2025) A Hybrid Deep Learning and Machine Learning Model for Predicting Student Academic Performance in Regular Basic Education. *Expert Systems with Applications*, **259**, Article 124123.
- [7] International Conference on Education Technology and Computers (2023) A Bibliometric Analysis of Machine Learning in Education: Research Themes and Emerging Trends. *Proceedings of the 2023 15th International Conference on Education Technology and Computers*, Barcelona, 26-28 September 2023, 334-341.
- [8] Khine, M.S. (2024) Machine Learning in Educational Sciences: Predictions and Applications. Springer. <https://doi.org/10.1007/978-981-99-9379-6>