基于 LightGBM 的大数据学业预警模型构建与评估

杜　平

广东建设职业技术学院 广东 广州 510440

【摘　要】：本研究基于多源学生数据构建了面向四级学业风险等级的分类与回归模型。实验结果表明，LightGBM 在回归评估中决定系数 R^2 为 0.6461，平均绝对误差为 0.1579，均方误差为 0.2146，展现出较高的数值预测精度。在分类任务中，模型整体准确率达 86.23%，对正常学生的识别召回率高达 98.75%。计算效率方面，LightGBM 训练耗时仅 0.3606 秒，单样本预测时间低至 0.0162 毫秒，能够满足大规模在线学业预警的实时性需求。研究表明，LightGBM 凭借其梯度提升框架与算法优化，在分类准确率、误差控制及计算效率上表现优异，是构建学情大数据预警系统的理想模型。

【关键词】：LightGBM；学情大数据；学业预警；梯度提升；分类预测；计算效率

DOI:10.12417/2982-3803.26.03.012

1 引言

随着教育信息化的发展，高校积累了海量的学情大数据，如何从中精准识别学业风险学生并实施及时干预，已成为教育数据挖掘的重要课题。机器学习方法近年来被逐步引入学业预警领域，其中集成学习模型凭借优异的泛化能力表现突出。LightGBM 作为一种基于梯度提升框架的优化模型，采用直方图算法与叶子生长策略，在大规模数据处理、训练效率及预测精度上具有显著优势，已在成绩预测等教育场景中展现出良好性能。然而，现有研究缺乏对 LightGBM 在学情大数据预警任务中的系统性性能评估[1]。

为此，本研究聚焦于 LightGBM 在学业预警中的建模与评估，基于多源学生数据，构建面向四级学业风险等级的预测模型，系统分析其在回归与分类任务中的预测精度、误差控制及计算效率。

2 理论基础

2.1 学业预警的传统研究方法

早期学业预警主要基于统计学方法与专家经验，常见做法包括：（1）描述性统计分析，如成绩分布、挂科率计算；（2）逻辑回归与判别分析，利用线性模型拟合学业表现与影响因素的关系；（3）指标体系构建与权重赋权，通过层次分析法、德尔菲法等确定多维度指标权重，形成综合评估分。这些方法虽具有一定解释性，但往往依赖主观假设、难以捕捉复杂非线性关系，且多为“事后归因”，预警时效有限。

2.2 机器学习在学业预警中的应用

随着教育大数据的积累与计算能力的提升，机器学习方法因其强大的模式识别与预测能力被逐步引入学业预警研究。当前应用主要涵盖三个方向：（1）分类预测，常用算法包括支持向量机、决策树、朴素贝叶斯以及集成学习模型；（2）特征重要性分析，通过模型输出识别影响学业的关键因素，如学习行为、心理状态、考勤记录等，为干预策略提供依据；（3）时序建模，利用循环神经网络、隐马尔可夫模型等方法分析学生行为的动态变化，实现更早的风险趋势预测。大量研究表明，机器学习模型能够从高维、非线性、不均衡的学情数据中提取有效信息，相较于传统方法，具有更高的预测精度与更强的泛化能力[2]。

2.3 LightGBM 模型及其优势

LightGBM（Light Gradient Boosting Machine）是一种基于梯度提升决策树的集成学习模型。相比于传统 GBDT，LightGBM 采用了两项核心技术优化：一是直方图算法，将连续特征离散化为固定数量的桶，大幅降低了计算复杂度与内存占用；二是叶子生长策略（Leaf-wise），每次从所有叶子节点中选择增益最大的节点进行分裂，在控制模型深度的同时显著提升精度。此外，LightGBM 原生支持类别特征处理、缺失值处理以及并行与分布式训练，使其在大规模学情数据场景下具有明显的效率优势。

3 研究方法

本研究采用数据驱动的研究范式，整合多源学情数据，基于 LightGBM 算法构建学业预警等级分类与回归预测模型，并

作者简介：杜平（1982--），女，工程师、信息系统项目管理师，华南理工大学硕士毕业，现任建筑信息学院大数据教研室主任，研究方向：人工智能、数据分析与应用。

基金项目：2023 年广东省高等职业教育教学质量与教学改革工程项目(课题编号:2023JG067)研究成果。

对模型性能进行系统评估。

3.1 数据来源与样本

本研究数据来源于两个渠道：一是面向在校大学生的匿名问卷调查，内容涵盖学习投入、学习习惯、心理状态（如焦虑、自我效能感）及环境感知（如师生关系、同伴支持）等多维度信息；二是学校教务管理系统导出的客观学业成绩记录，包括各科成绩、学分、平均学分绩点（GPA，满分 4.0）等关键数据。

3.2 变量说明

研究共选取并构建了 20 项特征变量，分为六大类别：（1）人口统计学特征，如性别、年级；（2）学习认知与行为特征，包括学习目标等级、课堂参与度、每周自主学习时长、作业完成情况等；（3）个人习惯特征，如睡眠时长、电子产品使用时间；（4）心理状态特征，包括学习焦虑程度、自我效能感评分；（5）环境支持特征，如家庭支持度、教师关注度；（6）客观学业特征，如上学期 GPA、已修学分、不及格门次等。预测目标为学业预警等级，该变量是根据 GPA 与不及格门次生成的四分类离散标签，用于标识学生的学业风险状态。

3.3 预警标签生成规则

为将连续的学业表现转化为适用于多分类监督学习的标签，同时兼顾教育管理实践中的常规标准，本研究依据平均学分绩点（GPA，满分 4.0）与不及格门次设定四级划分规则。最终规则如下：

等级 0（正常）：GPA≥3.3 且无不及格记录。

等级 1（关注）：GPA∈[2.8,3.3)或存在 1 门不及格。

等级 2（预警）：GPA∈[2.4,2.8)或存在 2 门不及格。

等级 3（高危）：GPA<2.4 或不及格门次≥3。

3.4 模型构建过程

3.4.1 数据预处理

原始数据经过初步整理后，依次执行清洗、转换与划分操作。对类别特征（如性别、年级、学习目标明确程度等）实施标签编码，将其映射为整数序列；连续型特征保持不变，直接输入模型。最后，将全部样本按照 8:2 的比例拆分为训练集与验证集，分层抽样保证各风险等级的比例与原集一致，随机数种子设为 42 以保障实验结果可复现[3]。

3.4.2 LightGBM 模型

LightGBM 在传统梯度提升框架上进行了多项优化。其核心思想是通过迭代训练一系列弱学习器（决策树），并以梯度下降方式最小化损失函数。对于训练样本（x_i, y_i），模型在第 t 轮的预测输出为：

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

其中f_t为当前轮的决策树。模型通过最小化包含正则项的目标函数来学习每棵树的结构：

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

其中为 l 可微损失函数，$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \| \omega \|^2$为正则化项，用于控制模型复杂度。

本研究使用 LightGBM 的 multiclass 目标函数，设置类别数为 4，并采用 multi_logloss 作为评估指标进行模型训练与优化[4]。

4 实验结果与分析

4.1 回归性能分析

将预警等级视为连续序数值（0、1、2、3），对 LightGBM 的回归预测能力进行评估。在测试集上，模型的决定系数（R^2）为 0.6461，表明模型能够解释约 64.6%的学业风险等级方差；解释方差分数为 0.6658，进一步验证了模型对数据变异的捕捉能力。在误差控制方面，平均绝对误差（MAE）为 0.1579，均方误差（MSE）为 0.2146。较低的误差值说明 LightGBM 在预测学生具体风险等级时具有较高的数值精度，能够为后续干预提供可靠参考。

4.2 分类性能分析

4.2.1 整体分类指标

LightGBM 在四级预警分类任务中的整体准确率达到 86.23%，表明模型能够正确识别绝大多数学生的风险等级。宏平均 F1 分数为 0.7337，综合反映了模型在各类别上的平衡表现。为排除随机一致性的影响，计算 Cohen‘s Kappa 系数为 0.7141，达到“高度一致”水平；马修斯相关系数（MCC）为 0.7224，进一步确认了模型在多分类场景下的强辨别能力。

4.2.2 混淆矩阵分析

模型对“正常”类别（等级 0）的识别极为精准，召回率达 98.75%，仅 1.25%的正常学生被误判为其他等级，有效避免了不必要的“误报警”。对于“关注”与“预警”两类中等风险学生，模型表现出一定的混淆，但主要误判发生在相邻等级之间，符合实际预警场景中边界模糊的特点。对于“高危”类别（等级 3），召回率为 50.00%，模型将另一半高危学生主要误判为“预警”等级。考虑到“预警”学生本身都是学校重点筛查与干预对象，通过后期的干预可以弥补缺陷，此类误判在

实际帮扶流程中仍能有效触发关注，具备较高的可用性[5]。图 1 展示了 LightGBM 在测试集上的混淆矩阵。

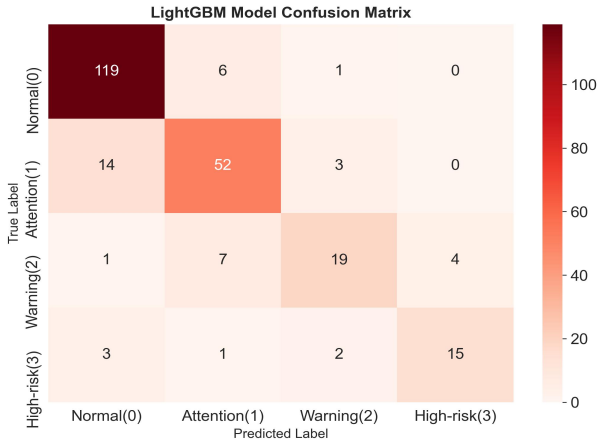

图 1 LightGBM 模型混淆矩阵

4.3 结果讨论

综合上述实验结果，LightGBM 在学业预警任务中展现出以下特点：第一，整体分类准确率高（86.23%），且 Kappa 系数与 MCC 均超过 0.71，表明模型预测与真实标签具有强一致性；第二，对“正常”学生的识别近乎完美，能够有效控制误报率，减轻教育管理者的无效工作负担；第三，对“高危”学生的召回率存在不足（50%），但误判主要流向相邻的“预警”等级，后续辅导员会对预警和高危的学生都进行逐个谈话和进一步干预，具有实践可接受性。计算效率方面，LightGBM 的训练与预测速度极快，非常适合部署于实时预警系统。后续可尝试引入代价敏感学习或异常检测技术，进一步提升对极端风险样本的敏感性。

5 讨论

LightGBM 在学业预警任务中展现出优异的分类精度（86.23%）与计算效率（单样本预测 0.016 毫秒），其梯度提升机制与叶子生长策略能够有效捕捉学情数据中的非线性特征交互，且预测误差控制在 0.16 个等级以内，具备精细区分相邻风险层级的能力。模型对“高危”学生的召回率仅为 50%，但混淆矩阵显示误判主要流向相邻的“预警”等级；在实际帮扶体系中，预警学生本身已是学校主动干预的对象，因此此类误判仍能使高风险个体进入关怀流程，具有可接受的容错性。

然而，研究数据源自单一院校、样本量有限，缺乏细粒度的课堂行为日志，且高危类别样本不足导致召回偏低。未来可引入代价敏感学习、过采样或异常检测技术，并融合多模态时序数据以提升对极端风险的识别能力。

6 结论与展望

本研究系统评估了 LightGBM 在学业预警任务中的性能。实验表明，LightGBM 分类准确率达 86.23%，平均绝对误差仅 0.1579，训练耗时 0.36 秒，单样本预测 0.016 毫秒，兼顾高精度与高效率，适合大规模实时预警系统。

未来，一是引入时序行为数据，构建动态预警模型；二是融合多模态信息，挖掘深层风险因子；三是采用代价敏感学习或过采样技术提升高危样本识别能力；四是开展跨校大样本验证，推动“预警-干预-评估”闭环落地。

参考文献：

[1] 王计生,徐多勇,唐莉,等.基于大数据的高校学生心理危机智能预警模型构建[J].成都医学院学报,2024,19(1):111-115.

[2] 任慧娜.基于校园行为数据挖掘的学生学业预警模型构建分析[J].信息系统工程,2026(3):22-25.

[3] 刘发稳,方芳,孙土土,谢显杰.基于群智感知数据的高职院校学业预警系统设计与实现[J].电脑知识与技术,2026(5):64-67.

[4] 舒仕文.LightGBM 模型及其应用[J].信息记录材料,2022(007):023.

[5] 李娟,郭蕊,郑天悦.基于动态多维学情画像的学生学业预警与干预系统设计[J].北京工业职业技术学院学报,2026,25(1):19-24.