

A Parallel Integrated Learning Model for Integrating Multi-Source Heterogeneous Data in the Study of Academic Risk Prediction for Vocational College Students

Jie Li Hongfang Cheng

School of Information and Artificial Intelligence, Wuhu Vocational and Technical University, Wuhu, Anhui, 241006, China

Abstract

To accurately predict students' dropout risks and assist vocational colleges in implementing early learning risk interventions, this study developed a student learning risk prediction model. The dataset underwent missing value imputation, normalization, one-hot encoding, and SMOTE-based category imbalance correction. The predictive performance of logistic regression, decision trees, Gradient Boosting, and Bagging models was compared. Experimental results demonstrated that the Bagging ensemble learning model achieved the best performance on the test set, with an accuracy of 99.62% and an F1-score of 85.33%. SHAP analysis identified mental health scores, GPA, and attendance rate as the three key predictors of student dropout. The proposed model exhibits high accuracy and interpretability, providing a scientific foundation for vocational colleges to establish precise early warning mechanisms and tailored intervention strategies.

Keywords

academic risk; multi-source heterogeneous data; integrated learning; vocational college students

融合多源异构数据的并行集成学习模型在高职学生学业风险预测研究

李杰 程鸿芳

芜湖职业技术大学信息与人工智能学院，中国·安徽 芜湖 241006

摘要

为准确预警学生的辍学风险，辅助高职院校开展早期学习风险干预。本研究通过构建学生学习风险预测模型，采取对数据集进行缺失值填充、标准化、独热编码与 SMOTE 类别不平衡处理，对比了逻辑回归模型、决策树、Gradient Boosting、Bagging模型的预测效果。实验结果显示，Bagging集成学习模型在测试集表现最优，准确率为99.62%、F1-Score为85.33%。SHAP技术分析表明，心理健康评分、GPA、出勤率是高职学生辍学的三大关键特征。研究构建的模型具备高准确率和可解释性，为高职院校构建精准预警机制与差异化干预方案提供了科学依据。

关键词

学业风险；多源异构数据；集成学习；高职学生

【基金项目】2025年度校级“教学质量与教学改革工程”

项目（项目编号：2025jyzd01）；2026年度校级“教学质量与教学改革工程”项目（项目编号：2026hxkc03）；2024年度高等学校省级质量工程项目（项目编号：2024wjdd001）；2024年度高等学校省级质量工程项目（项目编号：2024wjkc001）；2024年度高等学校省级质量工程项目（项目编号：2024jxzyk003）；2023年度省级一般项目——基于知识图谱的《数据结构》课程建设（项目编号：2023jyxm1311）；省级中青年骨干教师培育项目。

【作者简介】李杰（1996-），男，中国安徽合肥人，硕士，讲师，从事人工智能、数据挖掘研究。

1 引言

在教育领域，学生辍学、学习效果差异和个性化教学需求等问题一直备受关注，学生辍学问题在全球教育系统中普遍存在，尤其在高等职业院校中更为突出。辍学问题不仅影响学生的个人发展，还会对学校和社会资源造成浪费，加剧社会不平等。传统教学模式下，教师难以及时识别高风险学生，导致干预措施滞后。随着在线教育平台的普及和学习管理系统的广泛应用，教育领域正经历一场由数据驱动的变革。教育数据挖掘通过分析学习者的行为数据、课程交互记录及成绩等多维度信息，揭示学习规律并优化教学策略。

目前，现有的教育数据挖掘研究多聚焦于普通高校或基础教育领域，针对高等职业院校这一群体的数据挖掘仍较

少,高职院校学生中农村生源占比高、实践课程权重高、校企合作频繁等特性,要求研究者采用更具适应性的研究方法进行研究。因此,亟需针对高职院校学生的特点,开发具有高泛化能力的预测模型,以满足实际教育需求。针对上述问题,本文从教育理论出发,结合集成学习与可解释性分析方法,提出一种学生辍学风险预测框架,实现“预测—解释—干预”的闭环支持。

2 相关工作

2.1 学生学业风险预测中的机器学习应用

罗杨洋等 [1] 以混合课程为场景,对比五种机器学习算法,发现随机森林模型在预测成绩时准确率达 74.7%,并强调模型结构的教育意义可帮助教师理解学生行为与学习成绩之间的映射关系。该研究关于模型可解释性的讨论为后续研究提供了重要启示,模型的可解释性会影响教师的干预决策。Kord 等 [2] 提出结合预测与推荐的双阶段模型,使用支持向量机模型达到 78.04% 的分类准确率,表明了学生的课程数据对于提升预测精度中的作用。然而,研究重点在于验证双阶段模型的可行性与预测精度,而非重点探讨不同模型在计算效率和模型表达力上的理论边界。

2.2 多源异构数据的融合与特征工程

学业风险的形成是认知行为、学习过程、社会环境等多因素耦合的结果,单一的数据来源难以全面表现学生的学习状态。付玲毓等 [3] 利用真实慕课数据,将学生聚类为不同群体,发现在线时长和小测成绩是最终成绩的关键预测因素,为慕课教学的早期干预提供依据。该研究揭示了学习的过程性数据与学业成绩之间的关联程度,但尚未涉及更为丰富的多源异构数据源的建模预测。许惠惠 [4] 通过合并综合特征,构建多层感知机模型,为了解学生的思想政治情况提供数据支撑,该研究在数据层面整合了学生的基本信息、课程表现以及人格特质,但在模型层面仍属于异质集成,未充分利用同质集成中基学习器的优势。

2.3 模型可解释性与教育决策支持

针对学生学业风险预测的实用价值不仅依赖于预测精度,更取决于其对教学决策的可解释性支持。Lou 等 [5] 比较广义线性回归、决策树和随机森林回归,指出传统统计方法在特定教育场景中优于机器学习模型,强调了模型选择需结合数据特性,该研究体现了不同模型在追求预测精度的同时,需要考虑模型的教育语义可解释性。Yagci 等 [6] 以土耳其高校数据为基础,证明中期考试成绩可有效预测期末成绩,体现了 EDM 在早期预警系统中的实用性。与使用多维复杂数据的研究不同,该研究采用了极简的特征集,旨在论证利用有限数据来实现早期学业预警的实用性与可行性。

综上所述,多源数据融合的重要性在研究中已被认可。

然而,现有研究仍存在不足:针对高职院校的学生群体的研究相对较少,高职教育重点强调学生的技能培养与实训操练,其学生的数据特征与普通高等教育存在显著差异,且多数研究停留在“黑箱”预测层面,未能构建从风险识别到干预建议的可解释链路。

因此,本研究立足高职教育场景,构建融合多源异构数据的 Bagging 集成学习模型,通过 SHAP 等可解释性技术揭示高职院校学生学业风险的关键因素,为学校实施有效的教学干预提供理论与方法支撑。

3 研究方法

3.1 数据来源

本研究使用在线课程学生参与数据集,该数据集来自 kaggle,数据集中包含 1,000 条高等职业院校的学生数据,23 个数值型和类别型字段,包括:人口统计信息、学业表现、行为特征等信息,设数据集为

$$D = \{(x_i, y_i)\}_{i=1}^N$$

其中, x_i 表示第 i 个学生的特征向量, $y_i \in \{0,1\}$ 表示其辍学风险标签。有辍学风险的为 1,没有辍学风险的为 0。

3.2 数据处理

为减少数据泄露风险,保证模型评估无偏,本研究在对数据集进行处理前,将数据集划分成训练集和测试集,并采用分层抽样确保训练集和测试集中学生辍学率一致。其中,训练集用于学习数据特征,测试集用于评价模型性能。对数据中的缺失值进行处理,采用中位数填充连续型数值特征,减少极值影响,然后对数据进行标准化处理;采用众数填充类别型特征,减少偏差引入。针对异常值,使用缩尾处理将超出分位数范围的极端值替换成设定的边界值。针对有序变量,保留等级信息,对无序变量进行独热编码,不改变原始数据类别分布。针对数据集中,辍学样本与非辍学样本分布不均的问题,在模型的训练阶段,采取 SMOTE 过采样方法对辍学样本数据进行均衡化处理,提升模型对少数类别的识别能力。

3.3 模型构建

为精准识别高风险的学生,本研究选取两个基础模型:逻辑回归、决策树,两个集成学习模型:Gradient Boosting、Bagging 进行预测性能对比,其中,Gradient Boosting 属于串行集成学习的代表算法,Bagging 是经典的并行式集成学习方法。将准确率、F1-Score 值作为模型优劣的衡量指标,评估模型性能。准确率表示能够预测大多数正常就读学生的识别能力,但是当预测样本较少时,指标识别能力失真;F1-Score 值兼顾对辍学学生的识别能力以及预警精准度,评价维度更全面。各模型在测试集中的表现如图 1 所示。

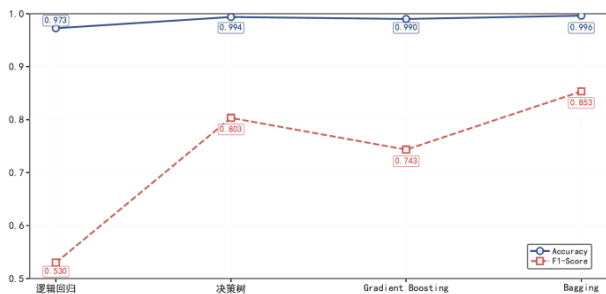


图1 各模型的准确率和F1-Score 比较图

4 实验结果与分析

从图1的实验结果可以看出,逻辑回归的准确率为97.25%,但是其F1-Score 值偏低,只有53%,不适合用于教育预测场景。Bagging 算法的性能优于基础的模型,也高于串行集成学习算法的 Gradient Boosting,预测效果最好。其中, Bagging 集成学习模型的准确率为99.62%,比基础模型决策树高2.37%。Bagging 集成学习模型的F1-Score 值为85.33%,比基础模型决策树高5%。

本研究中,测试集中的正样本数量较少,属于小样本不平衡场景。因此,研究采用5折交叉验证对模型的性能进行验证。结果表明, Bagging 集成学习模型在不同的数据划分情况下,模型效果依然可以保持稳定的预测效果,说明模型具有较好的稳健性和泛化潜力。

考虑到教育场景中对模型的可解释性,在保证模型性能的基础上,本研究采用 SHAP 方法对 Bagging 集成学习模型进行特征贡献度计算,如图2所示。SHAP 值显示“心理健康评分”对模型预测学生辍学贡献度最高,表明在校学生心理健康是需要学校教师重点关注的因素,可以采用问卷、访谈等方式了解学生的心理健康水平,并及时实施干预手段。

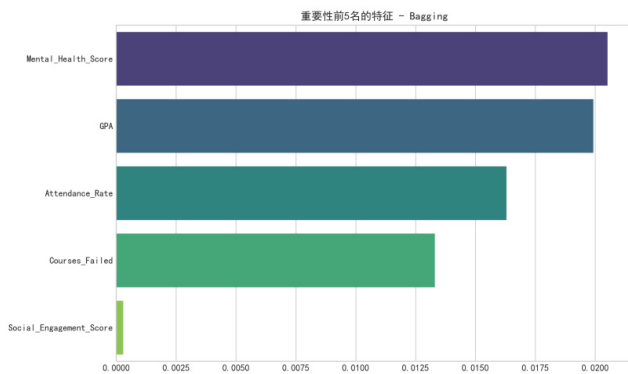


图2 重要性前5名的特征信息

模型识别出的 GPA、出勤率、未通过的课程数等因素对预测学生辍学也有重要影响,表明学业成绩会影响学生学

习的动力,进而导致学生产生辍学动机;出勤率可以体现学生对于课程学习的重视程度,低出勤率的学生,其学习热情较低,需要教师重点关注;未通过的课程数会让学生对专业学习感到压力,从而产生消极心理,因而更需要辅导员及时识别异常变化并介入支持。

5 结论与展望

本研究基于“预警学生辍学风险从而合理实施教学干预”的研究问题,通过构建 Bagging 集成学习模型实现对高职院校学生的学业风险预测,该模型在测试集中表现出高准确率和较好的稳定性。不同于传统的“黑盒”预测,研究通过 SHAP 技术实现了模型决策的透明化,明确了心理健康评分、GPA、出勤率为学生辍学的三大关键前兆,为高职院校实施教学干预提供了清晰的监测指标。

尽管 Bagging 集成学习模型的性能优秀,但是本研究仍存在一定的局限性。主要是体现在两个方面:一是教育数据集中样本数量相对有限,5折交叉验证证明了模型的稳定性,但是未来仍需要在更大规模、更多学校的数据集中验证模型的泛化性;二是当前的数据集中特征主要是学生的学业行为,在未来的研究中需要引入学习压力生理指标或者更多的在线学习行为日志指标,从而提升模型在早期预测的精准度。

参考文献

- [1] 罗杨洋, 韩锡斌. 混合课程学生成绩预测模型的可解释性探究[J]. 中国远程教育, 2022 (6): 46-55. DOI:10.13541/j.cnki.chinade.2022.06.006
- [2] Kord A, Aboelfetouh A, Shohieb S M. Academic course planning recommendation and students' performance prediction multi-modal based on educational data mining techniques[J]. Journal of Computing in Higher Education, 2025.DOI:10.1007/s12528-024-09426-0
- [3] 付玲毓, 贾积有. 基于教育数据挖掘的学习行为和学业成就分析——以北京大学某门慕课为例[J]. 教学研究, 2024, 47 (5): 1-8. DOI:10.13541/j.cnki.chinade.2024.05.001
- [4] 许惠惠. 基于多分支数据挖掘模型的大学生精准思政研究[J]. 智能计算机与应用, 2025, 15 (4): 115-120. DOI:10.20169/j.issn.2095-2163.24062702
- [5] Lou Y, Colvin K F. Performance prediction using educational data mining techniques: a comparative study[J]. Discover Education, 2025, 4: 112.DOI:10.1007/s44217-025-00502-w
- [6] Yagci M. Educational data mining: prediction of students' academic performance using machine learning algorithms[J]. Smart Learning Environments, 2022, 9 (1): 11.DOI:10.1186/s40561-022-00192-z