

高职院校学生学业预警模型构建研究

——基于 Logistic 回归与随机森林组合模型

罗卫红*, 李 燕, 张艺欣

江西航空职业技术学院 江西抚州

【摘要】当前国内学业预警研究多聚焦本科院校, 针对高职院校的量化建模较为匮乏, 普遍存在预警滞后、指标体系缺乏职业教育特色、模型可解释性与预测精度难以兼顾等问题。本文以高职院校学生为研究对象, 结合高职院校工学结合、以实训实习为主的办学特点, 融入实训成绩、岗位实习表现等特色指标, 构建基于 Logistic 回归与随机森林的组合学业预警模型。首先运用 Logistic 回归结合 LASSO 算法完成特征筛选, 明确学业风险核心影响因素; 再以随机森林作为主预测模型提升预警精度, 并通过 XGBoost、单一决策树开展对比验证。依托 AUC 值、准确率、召回率等指标完成模型评估, 最终划分红、黄、蓝三级学业风险预警等级, 配套制定分层分类干预策略, 形成从预警到干预再到反馈的闭环管理机制。

【关键词】高职院校; 学业预警; Logistic 回归; 随机森林; 组合模型

【收稿日期】2026 年 7 月 10 日

【出刊日期】2026 年 8 月 11 日

【DOI】10.12208/j.ssr.20260284

Research on the construction of an academic early warning model for vocational college students

——A hybrid model based on Logistic regression and random forest

WeiHong Luo*, Yan Li, Yixin Zhang

Jiangxi Aviation Vocational and Technical College, Fuzhou, Jiangxi

【Abstract】Most existing domestic studies on academic early warning focus on undergraduate universities, while quantitative modeling for vocational colleges is insufficient. Common problems include delayed early warning, an indicator system lacking features of vocational education, and difficulties in balancing model interpretability and prediction accuracy. Taking vocational college students as the research subjects, this paper conforms to the school-running characteristics of vocational colleges featuring work-integrated learning and prioritizing practical training and internships, incorporates characteristic indicators such as practical training scores and post internship performance, and constructs a hybrid academic early warning model based on Logistic Regression and Random Forest. Firstly, Logistic Regression combined with the LASSO algorithm is adopted for feature selection to identify the core influencing factors of academic risks. Random Forest is then used as the main prediction model to improve early warning accuracy, with XGBoost and a single decision tree applied for comparative verification. The model is evaluated by indicators including the AUC value, accuracy rate and recall rate. Finally, three tiers of academic risk warning levels (Red, Yellow and Blue) are divided, with hierarchical and classified intervention strategies formulated correspondingly, so as to form a closed-loop management mechanism covering early warning, intervention and feedback.

【Keywords】Vocational colleges; Academic early warning; Logistic Regression; Random Forest; Hybrid model

1 绪论

1.1 研究背景与意义

职业教育是我国国民教育体系与人力资源开发的

关键组成部分, 高职院校核心使命是培育技术技能型人才。当前, 高职招生规模不断扩大, 生源结构愈发多元, 学生普遍存在学习基础薄弱、学习动力不足、实训

*通讯作者: 罗卫红

适配能力差等问题,挂科、留级、退学等学业风险频发,既降低人才培养质量,也制约高职院校办学水平提升。

传统高职学生管理以事后处置为主,多在期末出现大面积挂科后才开展干预,时机滞后、整改成效有限^[1]。同时,现有学业预警研究多聚焦本科院校,指标体系侧重文化课成绩、课堂出勤,忽视实训课程、岗位实习等高职核心教学环节^[2]。此外,单一统计模型预测精度偏低,纯机器学习模型可解释性较差,无法兼顾学术严谨性与校园管理落地性,适配高职学情的专属预警体系较为匮乏。

基于此,本研究立足高职工学结合的办学特色,搭建适配高职学情的学业预警指标体系,融合 Logistic 回归与随机森林构建组合预测模型,兼顾统计模型的可解释性与机器学习模型的高精度优势。本研究可丰富职业教育量化研究思路,助力院校精准识别学业高危学生,推动学生管理从被动事后补救转向主动事前预判,有效降低学生学业风险,稳固技术技能人才培养质量^[3]。

1.2 国内外研究现状

国外学业预警研究历经三阶段:规则驱动阶段依托学分、绩点阈值事后预警,操作简便但滞后;传统统计阶段以 Logistic 回归为核心,可解释性佳,多维复杂数据下预测精度不足;机器学习阶段运用随机森林等算法,预测准确率超 94%,但模型为黑箱,难以溯源问题成因。国外职教虽重视学业管控,却因办学模式、生源与我国高职差异显著,研究成果无法直接照搬。国内相关研究近十年多聚焦本科院校,高职领域研究存在短板:部分仅探讨制度建设,缺少量化模型;部分直接套用本科指标,未纳入实训、岗位实习等高职核心要素。刘玲探讨高校学业预警机制优化路径,为预警等级划分提供参考^[4];张蕾围绕精准帮扶开展研究,完善学业预警的干预实施思路^[5];陈楚莎等针对学业困难群体搭建帮扶体系^[6];郭琪等、杨斌等分别采用不同机器学习算法开展学业预警建模实践^[7-8]。综上,国内高职学业预警研究尚存四大短板:研究重心偏向本科、指标体系缺失职教特色、预警介入时机偏晚、未构建完整预警干预闭环机制。

1.3 研究内容与方法

本文主要研究内容分为四部分:梳理学业预警研究演进历程与现存问题,明确高职学业预警的研究痛点与创新方向;结合高职办学特征构建包含基础指标与特色指标的预警指标体系,确定数据来源与样本范围;搭建 Logistic 回归和随机森林组合预警模型,完成

特征筛选、模型训练、多模型对比与效果评估;根据分析结果识别学业风险关键影响因素,划分三级预警等级,设计分层分类干预策略,构建预警、干预到反馈的闭环管理机制。研究方法包括文献研究法、实证研究法(选取某高职院校真实学生数据开展建模实证)和对比分析法(组合模型与单一 Logistic 回归、随机森林、XGBoost、决策树对比验证)。

1.4 研究创新点

本研究兼具多维创新:研究视角上聚焦高职教育细分领域,结合高职院校生源特征与人才培养模式开展实证建模,补齐现阶段量化预警研究短板;指标构建上结合工学结合特质,纳入实训课业成果、岗位实习成效、职业素养考评等专属指标,优化适配高职学情的预警指标维度;方法上整合 Logistic 回归与随机森林构建组合预测模型,兼容传统统计模型的可解释性与机器学习模型的精准度,弥补单一模型的应用局限;实践上突破重建模、轻落地的问题,依据预测结果划分差异化学业风险等级并配套分类干预策略,搭建全流程学业风险管理闭环,强化研究的落地应用价值。

2 相关概念与理论基础

2.1 核心概念界定

高职院校学业预警依托学生多维度学情数据,借助数据分析提前识别挂科、学业困难、退学风险学生并推送预警,是职业院校精细化学生管理关键手段。与本科院校不同,高职预警覆盖维度更全面,除理论课程外,同步纳入实训、岗位实习、职业素养考核,依托全过程数据实现风险前置研判,文中将多门课程挂科、实训考核不合格、岗位实习未通过统一判定为学业风险状态。Logistic 回归是经典二分类模型,适配二元结果研究,可量化各因素对学业风险的影响程度与显著性,常用于教育影响因素分析;随机森林基于决策树集成建模,通过多树投票完成预测,准确率高、过拟合风险低,还能输出变量重要性排序,更适配维度繁杂、存在非线性关系的高职工情数据。

2.2 理论基础

过程管理理论强调对事物发展全流程进行监控、干预与优化,而非仅关注最终结果。本文据此采集学生在校全周期的出勤、成绩、实训表现等过程性数据,实现全流程动态预警,做到早发现、早干预。教育大数据理论主张挖掘学生学习、生活、实践等多源数据的潜在价值,通过数据清洗与建模分析将零散数据转化为风险预判依据,推动学生管理从经验化向数据化转型。差异化管理理论要求针对不同群体制定个性化方案,不

同学业风险等级和成因的学生帮扶需求存在显著差异, 本文依据预警等级划分红、黄、蓝三类学生, 匹配分层干预策略, 提升帮扶工作的针对性与有效性。

3 研究设计与指标体系构建

3.1 预警指标体系构建

结合高职院校工学结合、知行合一的人才培养特质, 将影响因子划分为基础通用指标与高职特色指标两大类, 共计 18 项自变量, 涵盖学生个人禀赋、日常学习、在校行为、实践实训、家庭环境五大维度^[2]。基础通用指标适配全类型高校学情研判需求, 包含: 个人基础维度(性别、生源类型、入学成绩 3 项)、常规学习维度(理论课程平均分、课堂出勤率、迟到早退频次、课堂违纪频次 4 项)、在校综合表现维度(奖惩情况、课余自主学习时长、图书馆使用频次 3 项)、家庭环境维度(家庭所在地、家长文化程度 2 项)。高职特色指标贴合职业教育实践育人核心模式, 是区别于本科院校预警体系的专属指标, 包含实训课程平均分、单项实训考核合格率、实训出勤率、实训违纪频次、阶段性岗位实习评价得分、职业素养综合测评分数 6 项^[3]。

3.2 数据预处理

首先针对少量缺失值, 连续型指标采用均值填充法, 分类指标采用众数填充法, 同时剔除缺失率大于 20% 的无效样本。其次对入学成绩、课程得分、学习时长等连续型变量实施归一化处理, 消除数据量纲差异带来的建模干扰。最后对性别、生源类型、家庭所在地等分类特征进行独热编码转换, 将文本类分类数据转化为模型可识别的标准化数值数据。

4 组合预警模型构建与实证分析

4.1 整体建模思路

本文按照特征筛选、主模型构建、对比建模、模型评估的思路开展研究。先采用 LASSO (Least Absolute Shrinkage and Selection Operator) 结合 Logistic 回归筛选变量, 剔除弱相关指标, 提取学业风险核心影响因素, 明确各指标作用强度与方向, 保证模型可解释性。以筛选所得指标构建随机森林主预警模型, 提升预测精度。同步搭建单独 Logistic 回归、随机森林、XGBoost、决策树四类对照模型, 通过 AUC (Area Under the Curve)、准确率、召回率多维对比检验, 证实组合模型具备更优预测效果^[7]。

4.2 基于 Logistic 回归和 LASSO 的特征筛选

LASSO 算法可在回归过程中压缩无关变量系数至 0, 实现自动降维, 解决多变量共线性问题。将预处理后的全部指标纳入 Logistic-LASSO 模型, 结合显著性

检验 ($P < 0.05$) 筛选出 10 项核心影响指标: 理论课程平均分、课堂出勤率、实训课程平均分、实训出勤率、岗位实习评价得分、入学成绩、生源类型、课余学习时长、职业素养分数、违纪次数。从回归结果看, 理论课程成绩、实训成绩、出勤率、岗位实习得分越高, 学生出现学业风险的概率越低; 违纪次数越多、中职生源的学生, 风险概率显著上升, 印证了高职院校中实训学习与理论学习同等重要。

4.3 随机森林主模型构建

以 LASSO 筛选后的 10 项核心特征为输入变量, 学业风险二分类标签为输出变量搭建随机森林模型, 经网格搜索调优确定决策树数量、树深度、节点分裂数等核心参数, 完成模型训练。

4.4 多模型对比与效果评估

为验证组合模型预测效果, 本文采用统一测试集开展检验, 选用准确率、召回率、AUC 三项主流指标完成评价。其中准确率反映整体预测准确水平; 召回率为学业预警核心指标, 可降低高危学生漏判概率; AUC 用于衡量模型区分正常与风险学生的能力, 数值越趋近 1 性能越好。测试结果显示, 该组合模型准确率 93.12%、召回率 91.85%、AUC 达 0.942, 全部指标均优于对比模型。综上, 该融合模型融合两类算法优势, 既能依托 Logistic 回归解析学业风险诱因, 又借助随机森林实现精准预警, 适配高职院校学业风险预警实际需求。

5 学业预警等级划分与干预闭环机制

5.1 三级学业预警等级划分

基于组合模型输出的学业风险概率, 结合高职院校管理实际, 划定三级动态预警, 等级实时更新实现过程化管控。蓝色预警(风险概率 30%-50%)、黄色预警(50%-80%)、红色预警(80%以上)^[4]。

5.2 分层分类干预策略

结合预警等级、学业风险诱因构建分层差异化干预机制: 蓝色预警属预防性引导, 由任课教师、班委负责, 每周常态化沟通, 依托课堂观察、结对互助、主题班会柔性引导, 端正学习态度, 提前消解潜在学业隐患。黄色预警实施靶向重点帮扶, 辅导员与专业课、实训教师协同, 摸排学情、谈心, 研判风险根源再分类施策: 理论薄弱生组建补习小组, 实训短板学生一对一实操辅导, 厌学学生开展职业规划引导, 建立台账跟踪帮扶成效^[5]。红色预警启动刚性专项管控, 联动辅导员、班主任、家长、专业负责人四方协同, 每日追踪学习状态, 每周家校对接, 定制课业实训提升方案, 安排教师全天

候一对一辅导,从严管控考勤作息,视情况调整培养节奏,全方位处置学业危机^[1]。

5.3 预警-干预-反馈闭环管理机制

构建多环节联动学情全流程闭环管理体系。数据采集端由教务处、实训中心、学生处协同,常态化归集学生出勤、成绩、实训、违纪等全维度信息,动态更新数据库保障数据实时准确。系统依托组合模型每日测算,划分风险预警等级并分级推送管理人员。管理人员按预警层级开展分层帮扶,全程留存干预举措与学生动态台账。每周汇总干预数据优化模型参数阈值,每月召开学情专题会复盘机制。明确预警解除标准:学生连续两周风险概率低于 30%、课业实训回归常态,即可解除专项管控转入常规管理。体系打通学情数据、智能预警、校园管理链路,将学业预警转化为数字化、常态化日常管理手段。

6 研究结论、不足与展望

6.1 研究结论

本研究以高职在校生为对象,构建 Logistic 回归-随机森林组合预警模型及分级干预管理体系。主要结论:一是高职学业风险具有鲜明职教特征,实训成绩、岗位实习效果、实训出勤等实践指标是区分本科与高职学情风险的关键,预警体系须覆盖实践教学维度;二是组合模型准确率 93.12%、AUC 值 0.942,性能优于单一模型,可精准识别显性与隐性风险学生;三是按风险概率划分红、黄、蓝三级预警,匹配分层干预,形成预警—干预—反馈闭环,实现全周期动态管控;四是中职升学、违纪次数多、课余自主学习时间不足的学生风险较高,应作为重点帮扶群体。

6.2 研究不足与展望

本研究存在局限:仅选取单所高职院校,样本地域和专业覆盖面有限;指标体系侧重显性数据,未纳入心理健康、线上学习习惯等隐性变量;干预框架缺乏长效跟踪。未来可从三方面深化:拓展多区域、多类型高职

样本,优化模型参数;整合心理测评、线上学习轨迹、校园社交行为等多源数据,完善预警指标;依托实时数据流与 AI 技术开发智能预警平台,实现风险研判、精准帮扶、数据可视一体化,并建立长效跟踪机制,动态优化干预方案,构建数字化、智能化、常态化的高职学业风险管理体系^[6]。

参考文献

- [1] 郑仁顺.高校学业预警与帮扶实施,促进大学生学业发展[J]. 教育发展研究,2021(12).
- [2] 杜平,严琳.基于多源数据的高校学生学业危机预警指标体系构建研究[J]. 教育与研究,2026(1).
- [3] 吴越,周青,邓颖.精准思政视域下高校学业困难学生画像智能分析与预警模型构建研究[J]. 科技理论与实践,2025(8).
- [4] 刘玲.民办高校大学生学业预警机制优化及精准应对策略研究[J]. 当代教育与艺术,2026(6).
- [5] 张蕾.高校学生学业预警管理精准帮扶工作路径研究[J]. 发展教育学,2024(6).
- [6] 陈楚莎.新工科背景下学业困难学生的学业帮扶体系构建——以桂林电子科技大学信息与通信学院为例[J].2025年法治思想引领高校辅导员队伍专业化建设创新论坛论文集,2025(11).
- [7] 郭琪,田昊.基于支持向量机的机电逻辑控制类专业课学业预警模型构建与实践[J].计算机时代,2026(6).
- [8] 杨斌,张昱舟,张秀琴.基于 Apriori 算法的学业预警研究[J]. 网络安全与数据治理,2025(10).

版权声明: ©2026 作者与开放获取期刊研究中心(OAJRC)所有。本文章按照知识共享署名许可条款发表。

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



OPEN ACCESS