基于教育数据挖掘的大学生学业成绩预测与预警研究

田宇轩

中国石油大学（北京）克拉玛依校区 数据科学与大数据技术 24-1 班

新疆维吾尔族自治区 克拉玛依 834000

【摘　要】：高校在教学管理过程中积累了学生入学成绩、课程成绩和考核记录等数据。传统学业预警通常依据期末结果开展，难以在后续学期出现较大成绩波动前提供参考。本文采用公开高校学生学业数据，选取入学成绩及第一学期课程表现等特征，分别建立岭回归成绩预测模型和逻辑回归风险识别模型。实验以第二学期平均成绩作为预测目标，并以平均成绩低于 10 分作为低分风险的替代指标。对 4 023 条具有第二学期考核记录的样本按分层随机方法划分训练集和测试集。结果显示，增加第一学期数据后，岭回归模型的均方根误差由 4.1859 降至 2.5253；逻辑回归模型的风险识别 F1 值由 0 提高至 0.7485。与此同时，简单沿用第一学期平均成绩的基线方法在平均绝对误差方面仍优于本研究的岭回归模型，说明模型并非在所有指标上都具有优势。研究据此提出简单的学业风险分级提醒流程，为高校学业管理中的数据应用提供参考。

【关键词】：教育数据挖掘；大学生成绩；逻辑回归；岭回归；学业预警

DOI:10.12417/2705-0998.26.14.048

引言

大学生学业管理涉及课程成绩、修读进度和阶段考核等多方面信息。仅在学期末依据不及格记录开展预警，容易错过提前了解学生学习困难的机会。随着高校信息化建设推进，通过已有学业数据分析学生未来成绩已具备一定条件。机器学习算法能够利用历史样本中不同特征与结果之间的统计关系，为学业管理人员提供辅助判断。

现有研究方法较多，既包括复杂的集成算法，也包括逻辑回归、线性回归等基础模型。对于本科阶段的应用研究，选择清晰的数据集、合理的预测目标和可重复的实验过程，比追求算法复杂程度更重要。本文使用公开数据，将研究范围限定为“第一学期结束后预测第二学期平均成绩及低分风险”，重点说明数据处理、模型建立、模型训练与评价，并在此基础上设计基本的预警流程。

1 数据来源与实验设计

1.1 数据集来源

本文选取 UCI 机器学习数据库公开的“Predict Students'Dropout and Academic Success”数据集[1]。该数据集来源于葡萄牙高等教育机构，共包含 4 424 条学生记录和 36 个输入属性，内容包括学生入学成绩、原有学历成绩、课程修读与考核情况、第一学期和第二学期成绩等。公开数据已由发布方进行预处理，资料说明其没有常规缺失值[2]。

葡萄牙高校通常采用 0～20 分制，10 分为课程及格线[4]。需要注意，本文将“第二学期平均成绩低于 10 分”用作实验性低分风险指标，它不是单门课程是否不及格的直接记录，也不等同于学生最终是否退学。

1.2 研究样本与特征选择

为使预测目标具有明确的时间先后关系，模型只使用入学阶段信息和第一学期结束后已经形成的数据，不使用第二学期考核结果作为预测特征。第二学期平均成绩仅用于形成实际结果。为尽量排除尚未参加第二学期考核的记录，保留第二学期考核次数大于 0、且平均成绩位于 0～20 分范围的样本，共 4023 条。其中第二学期平均成绩低于 10 分的记录 469 条，其余 3554 条。

选取的特征分为两组。第一组为入学成绩和原有学历成绩，共 2 项；第二组在第一组基础上增加第一学期修课数、考核次数、通过课程数、平均成绩以及未参加考核课程数，共 7 项。通过比较两组模型结果，判断第一学期数据对后续学期预测是否有帮助。

表 1 主要研究变量

变量类别	变量名称	在模型中的作用
入学信息	入学成绩、原有学历成绩	基础预测特征
第一学期信息	修课数、考核次数、通过课程数、平均成绩、未考核课程数	增加的预测特征
第二学期信息	第二学期平均成绩	岭回归预测目标
风险标签	第二学期平均成绩低于 10 分记为 1，否则记为 0	逻辑回归预测目标

1.3 数据预处理与样本划分

首先依据变量含义筛选字段，检查是否存在不符合 0～20 分取值范围的第二学期成绩。随后将数值特征按 Z-score 方式标准化：

$$z=(x-\mu)/\sigma \quad (1)$$

其中，x 表示原始特征值，μ和σ分别表示该特征在训练集

中的均值和标准差。测试集只使用训练集计算得到的μ和σ，避免因提前接触测试集信息而影响评价结果。

数据按低分风险标签分层随机划分，80%用于训练，20%用于测试，随机种子为 42。最终训练集 3218 条，测试集 805 条，其中测试集包含 94 条低分风险记录。考虑本科研究的任务范围，本文采用固定的基础参数进行训练，不开展复杂的超参数搜索。

2 数学模型与训练方法

2.1 岭回归成绩预测模型

岭回归是在普通线性回归基础上增加 L2 正则化项的模型，适合以连续数值为预测目标的任务。设学生特征为 x1、x2、…、xm，第二学期平均成绩预测值为ŷ，则：

$$\hat{y}=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta mxm \quad (2)$$

训练过程通过调整参数β，使预测结果与实际成绩之间的误差尽量小。本文使用以下损失函数：

$$J(\beta)=1/(2n)\times\Sigma(yi-\hat{y}i)^2+\lambda/2\times\Sigma\beta j^2 \quad (3)$$

其中，n 为训练样本数，λ为正则化系数。实验固定λ=0.08，初始学习率为 0.045，采用全批量梯度下降更新参数，迭代 2500 轮。模型训练中不使用测试集成绩更新参数。

2.2 逻辑回归低分风险识别模型

逻辑回归用于二分类任务。将第二学期平均成绩低于 10 分标记为 1，其余标记为 0。模型先计算各项特征的线性组合，再通过 Sigmoid 函数输出低分风险分数：

$$P(y=1|X)=1/[1+e^{(-(\beta 0+\Sigma\beta jxj))}] \quad (4)$$

若输出结果大于或等于 0.5，则在本次实验中预测为低分风险。模型采用交叉熵损失函数并增加 L2 正则化项，固定λ=0.02，初始学习率为 0.18，使用全批量梯度下降训练 1200 轮。模型评价采用准确率、精确率、召回率、F1 值及 ROC-AUC。

2.3 评价指标

成绩预测采用平均绝对误差（MAE）、均方根误差（RMSE）和决定系数（R^2）评价。MAE 越小表示平均误差越小，RMSE 对较大预测误差更加敏感，R^2反映模型相对于测试集均值的拟合情况。

风险预测采用准确率、精确率、召回率和 F1 值。其中，召回率是实际低分风险样本中被正确识别的比例；F1 值用于综合评价精确率和召回率。由于低分风险样本少于正常样本，不能仅依靠准确率判断模型好坏。

3 模型训练结果与分析

3.1 低分风险识别结果

表 2 逻辑回归模型在测试集上的评价结果

特征组合	仅入学信息	加入第一学期信息
准确率	88.32%	94.91%
精确率	0	88.41%
召回率	0	64.89%
F1 值	0	0.7485
ROC-AUC	0.6456	0.9719

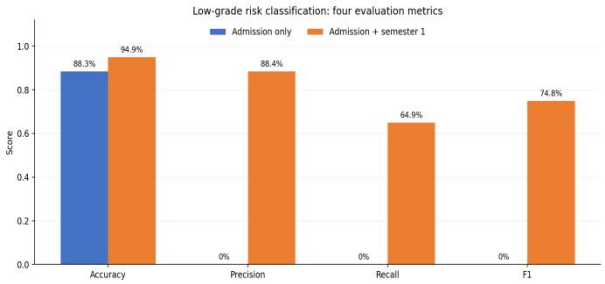

图 1 低分风险识别指标对比

从表 2 可见，模型仅使用入学信息时，将全部测试样本预测为非低分风险，因此虽然准确率达到 88.32%，但没有识别出任何真正的低分风险学生，召回率和 F1 值均为 0。加入第一学期信息后，模型能识别出部分风险学生，召回率为 64.89%，F1 值为 0.7485。

在 805 条测试记录中，加入第一学期特征的模型正确识别 61 条低分风险记录，将 8 条正常记录误判为风险，漏判 33 条风险记录，并正确识别 703 条正常记录。因此：

$$\text{准确率}=(61+703)/805=94.91\% \quad (5)$$

$$\text{召回率}=61/(61+33)=64.89\% \quad (6)$$

本研究还设置了一个简单对照规则：如果第一学期平均成绩低于 10 分，就预测第二学期存在低分风险。该规则在当前测试集上的准确率和 F1 值与逻辑回归模型相同。这说明逻辑回归模型虽然可以输出连续风险分数，但以 0.5 为阈值时，暂未在最终二分类结果上超过简单规则。因此不能仅凭复杂算法名称判断其应用价值。

3.2 第二学期平均成绩预测结果

表 3 岭回归模型与基础方法的测试结果（0～20 分制）

方法	MAE	RMSE	R^2
训练集均值预测	2.7313	4.2760	-0.0002
直接使用第一学期平均成绩	1.2738	2.8615	0.5521
岭回归：仅入学信息	2.6657	4.1859	0.0416

岭回归：加入第一学期信息	1.4447	2.5253	0.6512

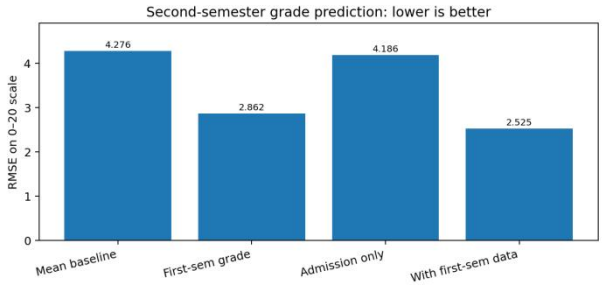

图 2 不同方法的成绩预测误差对比

由表 3 可以看出，使用入学信息预测第二学期成绩时，误差较大；加入第一学期信息后，岭回归的 MAE 下降至 1.4447，RMSE 下降至 2.5253，R^2提高至 0.6512。说明第一学期成绩与修课情况对于预测下一学期平均成绩具有明显帮助。

但是，直接沿用第一学期平均成绩的简单方法，其 MAE 为 1.2738，比岭回归的 1.4447 更低；岭回归的优势主要体现在 RMSE 和 R^2上。出现这种情况可能是由于两种指标对较大误差的敏感程度不同。实验说明，基础机器学习模型并不一定在所有指标上都优于简单预测方法，实际选择模型时应结合管理目标综合判断。

3.3 实验结果讨论

综合两项任务可以发现，第一学期学业信息比单纯入学信息更有预测价值。但研究也存在限制：第一，数据来自同一公开数据集，采用分层随机划分，没有进行跨年度验证；第二，实验风险指标是第二学期平均成绩低于 10 分，并不是“任意课程不及格”的真实记录；第三，数据没有按教学周提供完整的事件发生时间，因此实验只能验证第一学期结束后的预测，不能证明系统能够提前若干周发出预警；第四，模型没有接受真实学校的现场检验，也无法证明教学干预已经改善了学生成绩。

4 学业风险预警流程设计

4.1 风险信息生成

在第一学期结束后，教务人员导入学生入学成绩及第一学期修读信息。系统按照训练阶段保存的标准化参数处理输入数据，再利用逻辑回归模型计算第二学期低分风险分数。模型输出仅供辅助筛查，不能直接作为奖惩或学籍处理依据。

4.2 预警等级与处理方式

为便于教学管理，可初步设计三级提醒：风险分数低于 0.30 为常规关注；0.30 至不足 0.50 为需要进一步了解学习状态；大于或等于 0.50 为优先核查。上述阈值仅为管理流程设计示例，并未通过本次实验优化。教师应结合学生实际成绩、课程难度及沟通情况核实结果，避免单纯依靠算法判断。

4.3 反馈与复评

对于需要关注的学生，可由任课教师或辅导员了解学习困难，安排答疑或学习计划，并在后续阶段性考试后重新评估。学校应限制对个人学业数据的访问，保存数据使用记录。由于本文未开展实际部署，预警响应时间、真实干预效果等指标仍属于后续研究内容。

5 结论

本文基于 UCI 公开高校学业数据，采用岭回归和逻辑回归构建了第二学期平均成绩预测及低分风险识别模型，完成了样本筛选、特征标准化、模型训练、测试和结果分析。实验表明，在本数据集内，第一学期成绩及课程修读情况能够提高模型相对于仅使用入学信息时的预测性能；但与简单基线规则相比，模型优势因评价指标而异，尚不能直接证明复杂模型在真实高校预警场景中更有效。

本研究的主要价值在于展示一条操作简单、结果可核验的大学生学业数据挖掘流程。今后可在取得真实高校授权数据后，按照教学周划分数据开展时间外推检验，进一步研究预警阈值、误报控制和干预效果。

参考文献：

[1] Realinho V,Vieira Martins M,Machado J,Baptista L. 预测学生辍学与学业成功[数据集].UCI 机器学习存储库,2021.DOI:10.24432/C5MC89.https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

[2] Realinho V,Vieira Martins M,Machado J,Baptista L.预测学生辍学与学业成功[J].Data,2022,7(11):146.DOI:10.3390/data7110146.

[3] Pedregosa F,Varoquaux G,Gramfort A,等.Scikit-learn:Python 中的机器学习[J].机器学习研究杂志,2011,12:2825-2830.

[4] 葡萄牙高等教育总局.葡萄牙评分量表[EB/OL].https://www.dges.gov.pt/pt/pagina/escala-de-classificacao-portuguesa

[5] 黄焕,元帅.基于课程学习大数据的大学生学业预警机制及干预策略[J].高教论坛,2025,(11):89-93.