中山大學

SUN YAT-SEN UNIVERSITY

模式识别期末项目报告

学业困难学生识别

姓 名 陶宇卓、刘苇嘉、孙泽堃

学 号 22336216、22336155、22336210

学 院 计算机学院

专 业 计算机科学与技术

2025 年 7 月 3 日

目录

1 问题定义

随着在线教育平台的广泛应用，学生在虚拟学习环境中产生的行为数据（如视频观看、论坛互动、测验尝试等）为识别学业困难学生提供了新的可能性。传统教育干预往往滞后，无法及时帮助真正需要支持的学生。本研究利用模式识别分析技术，通过挖掘学生前几周的学习行为数据，构建预测模型来识别潜在的**学业困难**学生，实现早期预警和精准干预。

本研究旨在解决的核心问题是：如何基于学生早期学习行为数据，准确识别存在学业困难风险的学生。学业困难定义为：**课程不及格或最终辍学**。

2 数据集介绍

2.1 数据来源

实验使用 Open University Learning Analytics Dataset (OULAD) 数据集。
来源：英国开放大学 (Open University) 2013 - 2014 学年在线课程的真实匿名数据
特点：

- 覆盖 7 门不同学科课程
- 包含 22,000 + 学生记录
- 记录学习前 4 周的关键行为数据

学术价值：教育数据挖掘领域标杆数据集，被 ACM、IEEE 等顶级会议广泛采用

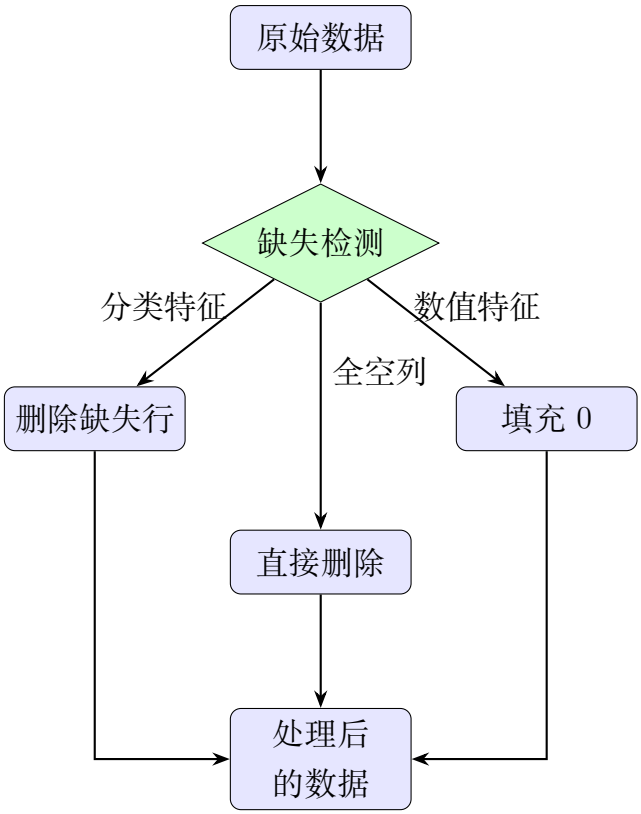
2.2 原始数据结构

表名	记录数	关键字段	描述
studentInfo	32,593	id_student, final_result	学生信息与最终成绩
studentVLE	10,655,280	sum_click, activity_type	学生-VLE 互动日志
vle	6,364	id_site, activity_type	VLE 资源元数据
courses	22	module_presentation_length	课程基本信息
studentRegistration	32,593	date_registration	学生注册信息
assessments	206	assessment_type	考核信息
studentAssessment	173,912	score	学生考核成绩

表 1: OULAD 数据集表结构

2.3 数据预处理

2.3.1 缺失值处理

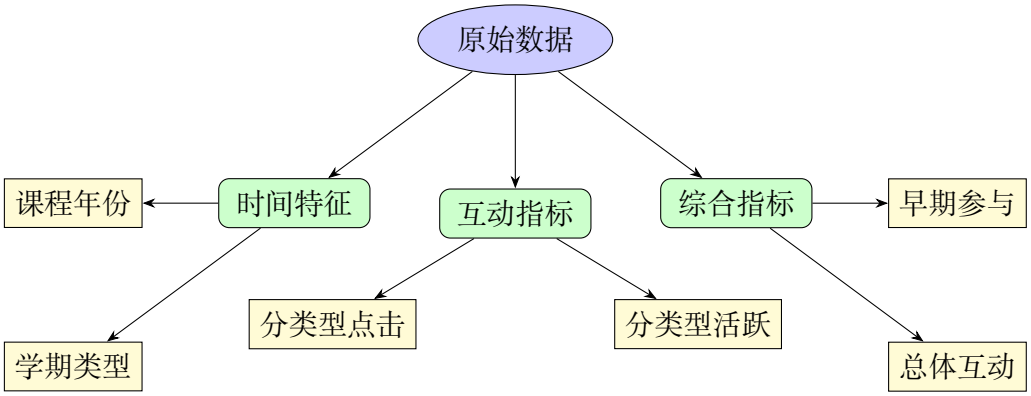


2.3.2 异常值处理

采用 98% 分位截断法，对超过阈值的极端异常值进行截断处理：

```
upper_limit = df[col].quantile(0.98)
df[col] = np.where(df[col] > upper_limit, upper_limit, df[col])
```

2.3.3 关键特征工程



基于” 早期识别学业困难学生” 的核心目标，对于虚拟学习环境（VLE）行为特征，通过多表关联构建互动指标，生成特征类型（举例如下，具体特征类型可见代码）：

特征类别	示例特征	计算逻辑
点击总量	total_clicks_forum	论坛总点击次数
活跃天数	active_days_quiz	参与测验的天数
综合指标	overall_total_clicks	所有活动类型点击总和
早期参与	early_engagement	前 4 周互动均值

表 2: 特征类别示例

2.4 目标变量

学业困难学生 (academic_risk) 的复合定义：

```
df['academic_risk'] = df['final_result'].apply(lambda x: 1 if x in
↪  ['Fail', 'Withdrawn'] else 0)
```

- **正例 (1)**：课程不及格 (Fail) 或中途辍学 (Withdrawn)
- **负例 (0)**：通过 (Pass) 或优秀 (Distinction)

3 算法原理

3.1 逻辑回归

逻辑回归（Logistic Regression）是一种用于分类问题的线性模型。其主要思想是通过 Sigmoid 函数将线性回归输出映射到 (0,1) 区间，表示事件发生的概率。模型形式为：

$$P(y=1|x) = \frac{1}{1+e^{-(w^T x+b)}}$$

通过极大似然估计方法进行参数训练，常用于二分类任务。

3.2 随机森林

随机森林（Random Forest）是一种集成学习方法，基于多个决策树构建。每棵树在训练时随机选取样本和特征，最终通过多数投票（分类）或平均（回归）得到预测结果。它具有较强的抗过拟合能力和良好的泛化性能。

3.3 KMeans 聚类

KMeans 是一种无监督学习算法，用于将数据划分为 K 个簇。算法通过迭代以下两个步骤进行：

- 分配步骤：将每个样本分配到最近的簇中心；
- 更新步骤：重新计算每个簇的中心为其成员点的均值。

直到簇中心收敛或达到最大迭代次数。

3.4 支持向量机

支持向量机（SVM）是一种二分类模型，其基本思想是寻找一个能够最大化分类间隔的超平面。对于非线性问题，可通过核函数将数据映射到高维空间以实现线性可分。其优化目标为：

$$\min \frac{1}{2}||w||^2 \quad \text{subject to } y_i(w^T x_i + b) \geq 1$$

3.5 决策树

决策树通过对特征进行条件判断构建模型，其核心在于选择最优特征进行分裂。常用的划分准则包括信息增益、信息增益率和基尼指数。模型可解释性强，但容易过拟合，通常结合剪枝策略提升泛化能力。

3.6 神经网络

神经网络由输入层、隐藏层和输出层构成。每层由多个神经元组成，使用激活函数（如 ReLU、Sigmoid）引入非线性。通过反向传播算法计算梯度，使用优化器（如 SGD 或 Adam）进行参数更新。适用于建模复杂非线性关系。

4 实验设置及评估标准

4.1 数据集划分

训练集：预处理后数据的 **80%**（25185/31482）
测试集：预处理后数据的 **20%**（6297/31482）
采用分层抽样保证分布一致：

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
↪  test_size=0.2, stratify=y, random_state=42)
```

使用 SMOTE 处理类别不平衡:

```
smote = SMOTE(random_state=42)
X_train, y_train = smote.fit_resample(X_train, y_train)
```

4.2 模型性能评估

评估指标	意义
准确率 (Accuracy)	衡量模型整体预测正确率，反映模型对正常学生和学业困难学生的综合识别能力
精确率 (Precision)	表示模型预测为”学业困难”的学生中实际困难学生的比例，评估干预资源的投放效率
召回率 (Recall)	反映模型识别实际学业困难学生的能力，衡量学业困难高风险学生的漏报率
F1 分数	综合精确率和召回率的平衡指标，评估模型在不平衡数据上的稳健性
ROC AUC	衡量模型区分正负样本的能力，评估模型的风险排序能力

表 3: 评估指标及其意义

4.3 SHAP 特征分析

SHAP（SHapley Additive exPlanations）是一种基于博弈论的特征解释方法，其核心原理是将模型预测值分解为各特征的贡献值，从而精确测量每个行为特征对风险预测的影响程度。在本次实验中我们使用了这种分析方法。

关键步骤:

1. **样本选择:** 从训练集中随机抽取 1,000 个样本（具体样本数量根据实际算法运行效果决定，以保证代表性的同时提高计算效率）
2. **值计算:** 使用 TreeSHAP 算法高效计算每个特征的 Shapley 值
3. **全局分析:**
   - 特征重要性排序（基于平均绝对 SHAP 值）
   - 特征效应方向分析（正负 SHAP 值分布）
4. **局部分析:**

- 个体预测解释（特定学生的风险因素分解）
- 典型样本案例分析（高低风险学生对比）

4.4 GridSearch 自动化调参

GridSearch（网格搜索）是一种系统地遍历多个参数组合来寻找模型最优超参数的方法。它通过对指定参数空间进行穷举搜索，并结合交叉验证评估模型性能，最终选出性能最佳的参数组合。在本次实验中，我们使用了 `GridSearchCV` 工具对部分模型进行了自动化调参。

关键步骤：

1. **参数空间定义：**根据模型的性质，设定待调节的参数组合范围。
2. **评分指标选择：**选择合适的评分标准作为参数优劣的评估依据。
3. **交叉验证：**采用 k 折交叉验证对每组参数组合进行评估，减少过拟合风险。
4. **最优参数选取：**根据平均交叉验证得分选择最佳参数组合，并用于后续模型训练和测试。

5 实验结果

5.1 整体性能概览

本实验评估了六种模型在“学业困难学生”识别任务中的性能表现，包括准确率、F1 分数、召回率、精确率、ROC-AUC 以及训练时间等多个指标。图 1 展示了各主要指标的对比，图 2 展示了整体综合评分情况。

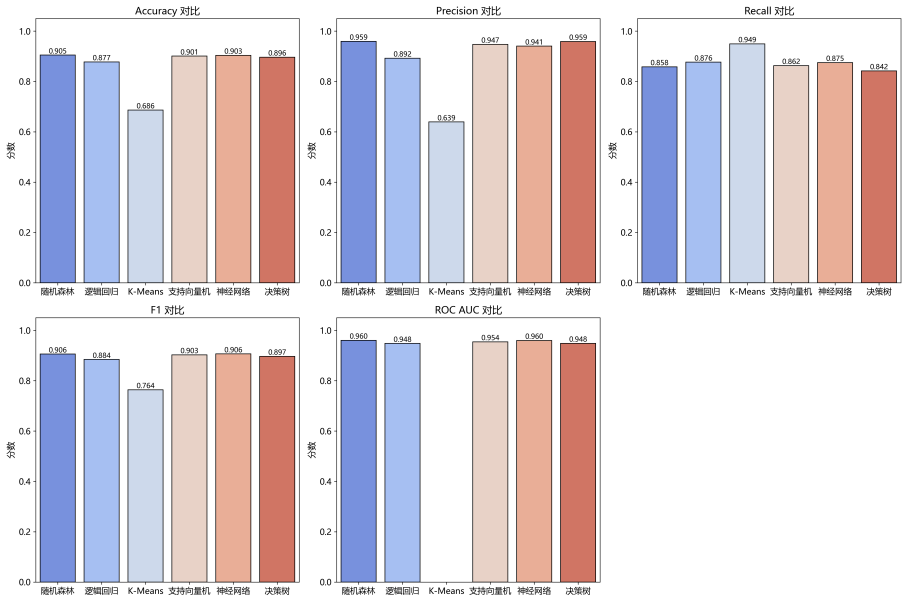


图 1: 各模型性能指标对比（Accuracy、F1、Recall、Precision、ROC-AUC）

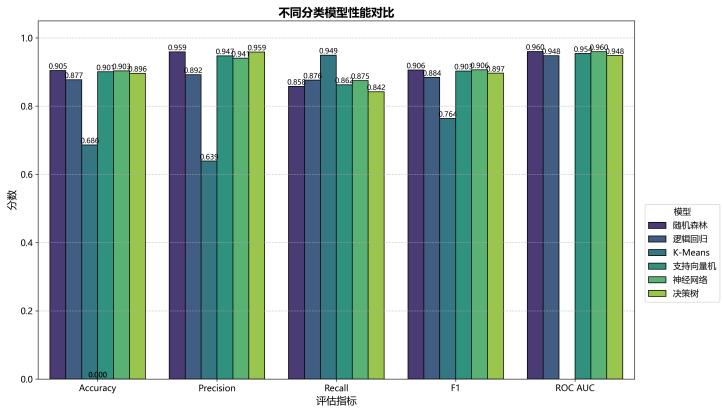


图 2: 各模型综合评分比较

表 4 总结了各模型的数值指标表现,其中随机森林综合得分最高 (综合得分: 0.9175):

表 4: 各模型主要性能指标对比

模型	准确率	F1 分数	Precision	Recall	ROC-AUC
KMeans	0.6862	0.7638	0.6390	**0.9492**	–
决策树	0.8961	0.8966	0.9587	0.8419	0.9481
支持向量机	0.9007	0.9028	0.9471	0.8624	0.9542
神经网络	0.9034	**0.9064**	0.9406	0.8746	0.9596
逻辑回归	0.8774	0.8843	0.8923	0.8764	0.9479
随机森林	**0.9046**	0.9058	**0.9591**	0.8580	**0.9602**

5.2 分类效果与混淆矩阵

为了直观展示各模型对“风险学生”的分类能力，图 3 展示了典型模型的混淆矩阵结果。从中可见，随机森林和神经网络在识别正类（风险学生）方面表现良好，误判较少。

5.3 特征重要性分析（SHAP）

我们使用 TreeSHAP 方法对模型的特征重要性进行了可解释性分析。图 4 展示了所有样本的 SHAP 值分布情况，图 6 对“活跃天数”这一关键特征的影响进行了深入分析，图 5 展示了平均特征重要性排序。

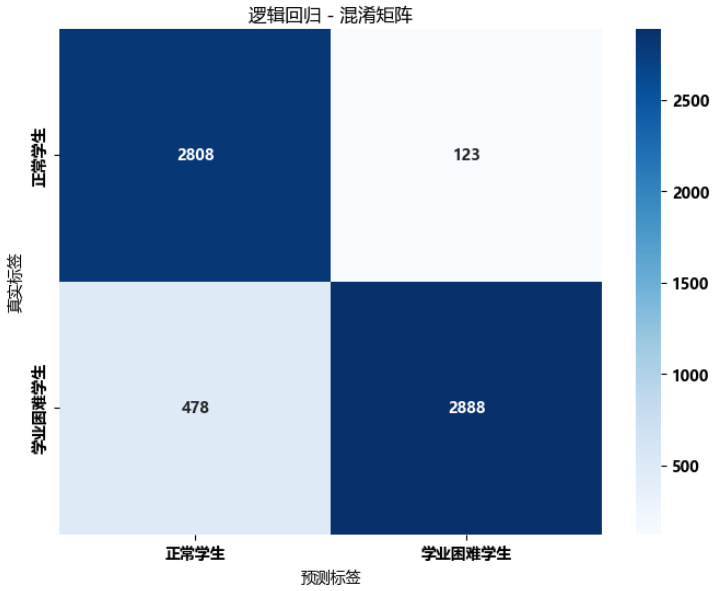


图 3: 随机森林的混淆矩阵

6 对比分析及结论

6.1 从多维度对比算法性能

- **准确性**：随机森林和神经网络表现最优，在多个指标均为九种算法中的最高值。
- **训练效率与资源开销**：逻辑回归与决策树训练最快，模型体积小。神经网络耗时最长但预测效果好（具体结果见附件）。
- **特征敏感度与可解释性**：逻辑回归与决策树可直接输出权重或规则，解释性强。神经网络表现较黑箱，需借助 SHAP 工具分析。
- **调参难度**：逻辑回归、决策树参数简单；SVM 和神经网络参数较多，需配合 GridSearch 等自动化工具调参。
- **鲁棒性与不平衡处理能力**：随机森林和神经网络在处理“风险学生”这种少数类上更鲁棒。KMeans 虽 Recall 高但 Precision 低，不适合实际部署。

6.2 不同模型对关键特征的判断是否一致

从 SHAP 分析中发现，各模型普遍认为以下四个特征在预测学业风险时最为关键：

- 活跃天数（overall_active_days）
- 主页访问次数（active_days_homepage）
- 总点击次数（overall_total_clicks）

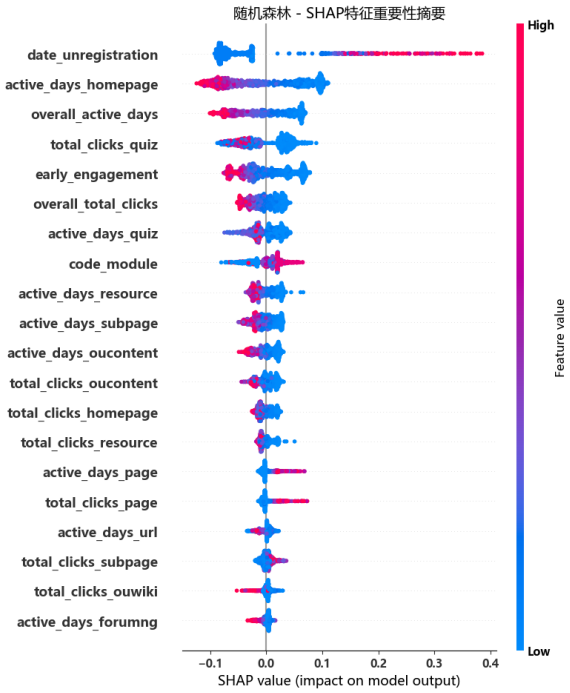


图 4: SHAP Summary Plot：各特征重要性及其影响方向

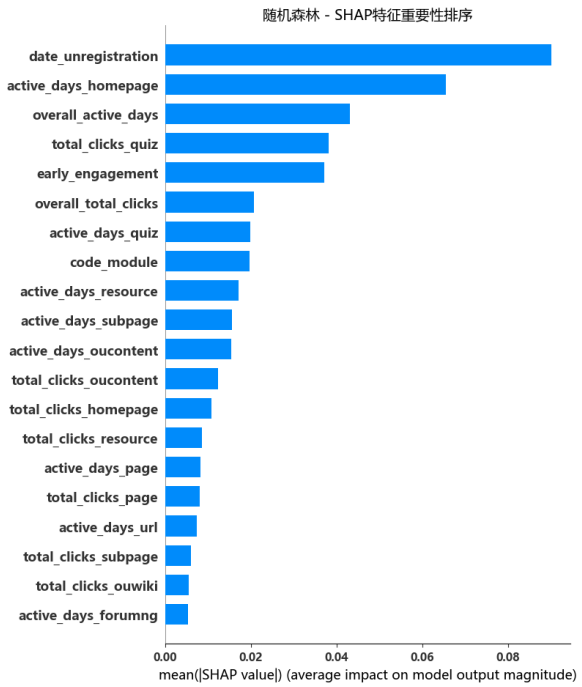


图 5: 特征重要性排序条形图（按平均 SHAP 值）

- 课程退课日期（date_unregistration）

模型间在主导特征上的判断高度一致，但对于部分边缘特征在不同模型中权重差异较大。

6.3 少数类识别能力对比分析

“风险学生”为少数类，对 Recall 要求更高。从实验结果看，KMeans Recall 虽高（0.9492），但 F1 分数偏低，预测准确性不稳定；随机森林与神经网络则在 Recall 与 Precision 之间取得了较好平衡，表现出较强的鲁棒性。这与它们分别采用了“集成学习”与“非线性表达能力强”的结构特点有关。

6.4 模型适用场景与局限性

- **随机森林**：适用于对预测准确性要求高、数据维度中等的场景，部署成本适中，鲁棒性强。
- **神经网络**：适合大数据场景，对特征之间的复杂非线性关系建模能力强，但可解释性差，训练耗时长。
- **逻辑回归**：适用于需要快速建模和解释的业务系统，简单高效。

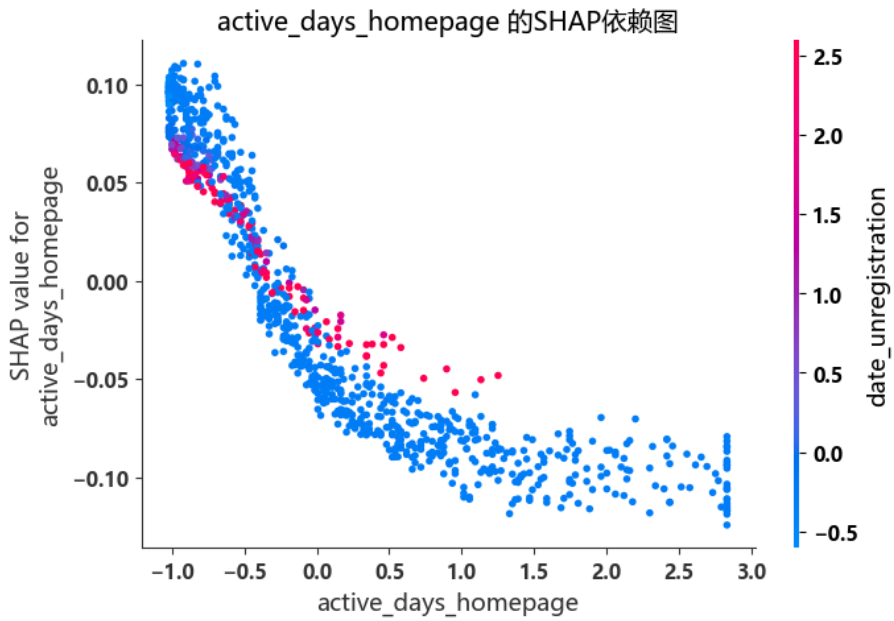


图 6: SHAP 依赖图：活跃天数与退课日期对风险预测的影响

- **决策树**：结构直观，适合入门和教学场景，但容易过拟合。
- **支持向量机**：对中小规模数据建模效果好，但调参困难。
- **KMeans**：不适用于监督分类，仅限于聚类和初步探索分析。

7 未来发展思考

未来研究可从以下几方面进一步提升学业风险预测系统的表现和实用性：

- **引入时间序列模型**：如 LSTM、Transformer 等结构建模学生行为序列，捕捉长期动态特征。
- **优化不平衡处理策略**：结合 SMOTE、聚类采样或代价敏感学习增强对少数类的识别能力。
- **模型压缩与部署**：探索模型剪枝、量化技术，使神经网络等复杂模型可应用于轻量级终端。
- **多模态数据融合**：整合日志、问卷、成绩等多维数据，构建更加全面的学生画像。
- **提升可解释性**：集成 SHAP、LIME 等方法，向师生用户提供透明、可信的个体预测解释。