北京航空航天大學
BEIHANG UNIVERSITY

数据科学导论期末大作业

高校学生辍学风险模型构建与分析

组员：李卓衡、高铭、何家天、郑慕函

2025 年 5 月

目录

一、 引言

在数字化浪潮推动下，高等教育正面临前所未有的变革，在线学习平台、混合式教学模式的普及打破了时空限制，为学生提供了更加灵活的学习路径。然而，与此同时，学生辍学率居高不下，已成为制约教育质量和公平的重要挑战。根据联合国教科文组织（UNESCO）发布的数据，全球高校平均辍学率约为 19%，在部分发展中国家甚至超过 30%。特别是在在线教育或技术类专业中，学生更容易因课程难度、缺乏互动或个人支持系统不足而中途放弃学业。

学生辍学不仅是教育体系内部的问题，也带来了深远的社会经济影响。以美国为例，据统计，高中辍学者的长期失业率高达 56%，其一生的平均收入损失超过 26 万美元。这种个体层面的损失在国家层面进一步积累，造成教育资源的浪费与社会成本的上升。此外，家庭经济状况、学术适应能力、心理健康和政策支持等多重因素相互交织，使得学生升学路径的预测与管理愈加复杂化。

在此背景下，教育数据挖掘（Educational Data Mining, EDM）与学习分析（Learning Analytics, LA）作为新兴的交叉研究领域，日益成为解决该问题的重要技术路径。通过整合学生在学期间的学术记录、行为数据、人口统计与社会经济变量，结合机器学习与可视化分析手段，研究者和教育管理者可以更早识别潜在的高风险学生群体，及时采取个性化干预措施。

本研究立足于欧洲高校十年间真实学生数据集，从学生个体特征、家庭背景到学习表现、宏观经济环境等六大维度出发，构建了完整的数据预处理、特征分析、模型预测与可视化流程。我们旨在实现以下三点目标：

1. 识别影响学生升学状态（毕业、在读、辍学）的关键因素，为教育实践提供理论支持；
2. 构建高准确性与可解释性的预测模型，提升教育管理的信息化与精准化水平；
3. 探讨非成绩因素对升学路径的影响机制，为教育公平与学生支持政策提供实证依据。

通过本研究，我们希望为高校教育治理提供更加智能化、科学化的决策工具，并推动教育资源向真正有需要的群体倾斜，实现教育公平与效率的双重提升。

二、 文献综述

高等教育中的学生辍学问题长期制约教育公平与资源配置效率，全球平均辍学率约为 19%，在发展中国家甚至超过 30%（UNESCO, 2021）。随着教育数据挖掘（EDM）与学习分析（LA）技术的发展，研究逐渐从经验归纳转向数据驱动的建模方法，并融合社会实证与政策分析，形成跨学科研究趋势。

早期研究多聚焦宏观社会因素。Kehm 等通过综述 44 项欧洲实证研究，提炼出九大影响辍学的核心要素，其中社会经济地位被认为是最稳定的预测因子[1]。相关研究表明，家庭收入每下降 10%，辍学风险上升至 3.2 倍，父母教育水平亦显著影响学生完成学业的概率[5]，印证了“资源剥夺理论”的解释力。

在方法方面，早期研究多采用逻辑回归或描述性统计。Atchley 等发现，技术类专业在线学习的完成率低于传统课程（93.3% vs 95.6%），但由于缺乏多源数据整合，难以揭示具体机制[2]。而 Daud[7]引入“家庭支出”“自雇状态”等特征，使模型准确率提升至 86%，证明多元变量可提升预测性能。

技术类专业的高辍学率尤为突出，原因包括课程难度高、实践压力大，以及相关行业不确定性高（如生物燃料产业收缩）[4]。这些结论与 Atchley 等的研究相符，后者指出金融课程完成率为 82.2%，而阅读类课程高达 98.2%。

在政策层面，欧盟委员会指出应通过包容性教育与学生中心化模式支持弱势群体[3]。目前研究普遍强调将模型识别出的高风险群体（如技术类学生、低参与度学生、社会经济弱势群体）与政策手段结合。例如，对技术类学生可加强实践辅导与职业规划教育；对低参与学生，可通过学习管理系统监测其跨平台活跃度并实施预警干预[6]。

尽管已有大量关于学生辍学预测的研究，但仍存在若干不足。首先，多数研究将问题简化为二分类任务，忽略“在读”状态的特殊性，可能引入噪声并影响预测效果。其次，传统模型对复杂非线性关系和类别型变量的处理能力有限，难以准确捕捉辍学风险因素。同时，现有研究多关注整体准确率，缺乏对少数类识别效果的分析，难以支撑实际干预需求。此外，模型解释性与特征影响分析常被忽视，限制了研究的实际应用价值。

本研究尝试回应上述不足：通过比较三分类与二分类策略，明确“在读”状态对预测的干扰，并重构任务提升性能；引入 CatBoost 和 LightGBM 等更先进的模型，并结合 SMOTE 与调参技术，提高了模型的适应性与少数类识别能力；同时系统分析了多类因素对辍学的影响，增强了研究的实用性与可解释性。

三、 数据集描述与预处理

3.1 数据集来源

我们选取 Student Dropout Analysis for School Education 数据集作为我们的数据科学分析数据集[8]。该数据集由波塔莱格雷理工学院创建，使用的数据源来自：(i)机构的学术管理系统(AMS)，(ii)机构的教学活动支持系统（内部开发，称为 PAE），(iii)高等教育总司(DGES)关于通过国家高等教育入学竞赛(CNAES)录取的年度数据，以及(iv)关于宏观经济数据的当代葡萄牙数据库(PORDATA)。数据集描述了真实世界中欧洲多个高校 2008/2009 学年（在欧洲高等教育博洛尼亚进程实施之后）至 2018/2019 学年期间注册学生的记录。该数据集同时被开源数据集网站 kaggle 记录，我们从 kaggle 中获取得到具体的数据集 csv 文件，并基于该数据集文件开展之后的分析研究任务。

3.2 数据集描述

该数据集包含 4424 条记录和 35 个属性，从 6 个数据维度记录了学生的升学信息。

(i) 人口统计学数据：包括婚姻状况、国籍、性别、年龄、是否为迁移人员、是否为国际学生 6 个属性

(ii) 社会经济状况数据：包括父亲学历、母亲学历、父亲职业、母亲职业、特殊教育需求、是否为债务人、学费缴费状态、是否为奖学金获得者 8 个属性

(iii) 宏观经济环境数据: 包括失业率、通货膨胀率、国家 GDP3 个属性

(iv) 入学注册时的学术数据: 包括入学申请方式、入学申请志愿、专业课程、出勤状况、先前学历 5 个属性

(v) 第一学期结束时的学术数据: 包括第一学期的已获学分课程、注册课程、已评估课程、已通过课程、课程成绩、未评估课程 6 个属性

(vi) 第二学期结束时的学术数据: 包括第二学期的已获学分课程、已注册课程、已评估课程、已通过课程、课程成绩、未评估课程 6 个属性

3.3 数据预处理

3.3.1 数据缺失情况

在数据预处理阶段，我们首先对原始数据集中的缺失值情况进行了系统性检查。我们使用 pandas 库的缺失值统计方法，对全部特征变量进行了缺失值计数。统计结果显示，在原始数据集中包括个人基本信息、家庭背景、学业表现、经济信息、宏观经济特征以及目标变量在内的所有特征均无缺失值，即缺失值数量均为 0。

这一结果说明所使用的数据源质量较高，在样本收集和处理阶段已经进行了较为完备的清洗和验证，避免了常见的缺失数据问题。因此，本研究可以直接进入异常值检测与数据清洗等步骤。

3.3.2 数据异常

为确保模型训练过程中不受到极端值的干扰，本文对数据集中的异常值进行了系统性检测。为了量化数值型变量中的潜在异常值，本研究计算了每个变量的 Z-score 值，并以阈值 ±3 作为判别标准。统计结果表明，大多数变量中异常值比例较低(小于 3%)，说明数据总体分布较为合理，未出现大规模的离群现象。尽管数据集中存在个别变量的轻度异常值，但整体而言，数据质量较为稳定，异常值比例控制在可接受范围之内。因此，后续建模阶段未对异常值进行删除处理，而是通过集成模型的鲁棒性予以缓解。

3.3.3 数据清洗

为进一步提升数据质量与建模效果，本文对原始数据进行了必要的清洗操作，包括字段命名标准化与目标变量的数值编码。

在目标变量 Target 的处理上，原始数据中该变量为字符串类型，包含三类取值：'Dropout'（辍学）、'Enrolled'（在读）以及 'Graduate'（已毕业）。为了便于模型进行监督学习并兼顾预测任务的语义解释性，本文将该变量映射为数值型变量，采用如下编码方式：'Dropout' → 0，'Enrolled' → 1，'Graduate' → 2。这种编码方式体现了对学生学业进程的**线性阶段性理解**：0 表示未完成学业（辍学），1 表示正在就读，2 表示顺利完成学业（毕业），有助于部分机器学习模型（尤其是树模型）理解状态之间的“相对距离”。

注意到原数据集中部分类型为 int32 的数值并无实际意义，仅为类型代号，因此我们将这些列改为 category 类型。对于数据集中的 34 个特征，我们的目标是研究它们与

目标变量的关系。因此我们选择将标签没有关联的特征从建模中删除，方法是对分类变量与因变量的相关性进行卡方独立检验。通过计算我们发现除了三个变量（'Nationality', 'International', 'Educational_special_needs'）的 p 值非常高（0.24、0.53、0.73）外，大多数 p 值接近于零，这表明这三个特征与标签之间没有统计学上显著的关联。我们将把这些变量排除在建模之外。

综上所述，通过字段修正、数据过滤、标签编码与变量清理等清洗操作，本文为后续建模工作提供了更加整洁、规范的数据基础。

四、 探索性数据分析（EDA）

EDA 的目的是使用统计图表和其他数据可视化工具来了解数据的主要特征，在这一步骤中，我们通过计算皮尔逊相关系数来衡量数值变量之间的线性关系强度。对于分类目标变量，我们首先将其进行数值编码，然后计算各特征与目标变量之间的相关系数。这一方法可以帮助我们发现特征间的多重共线性问题，为后续的特征工程和模型构建提供指导。

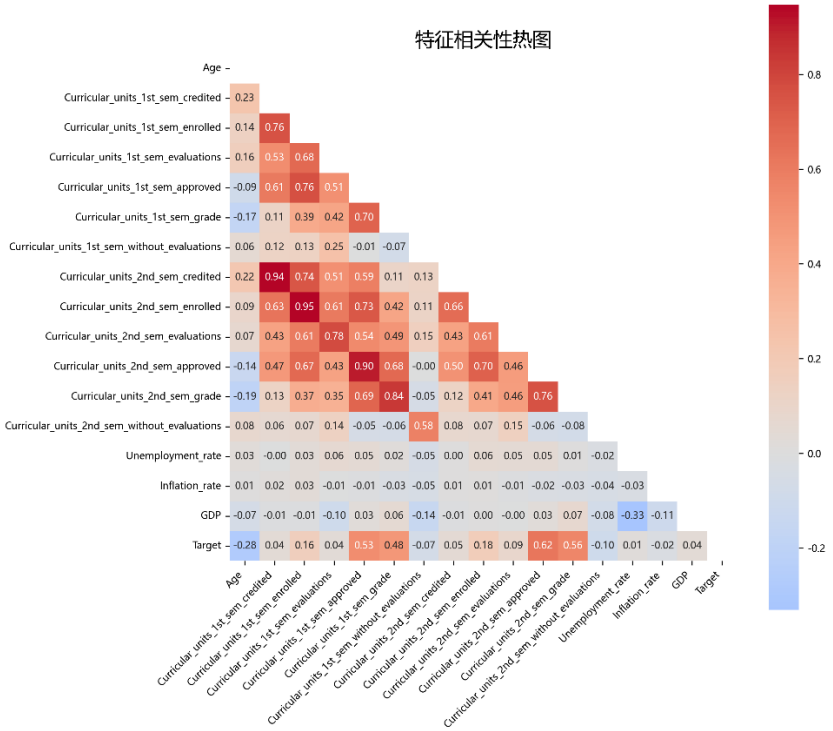

图 1 数值型变量相关性分析热图和相关性柱状图

我们计算了所有数值特征与 Target 的皮尔逊相关系数，并以热图的形式进行了可视化。从互相关热图中我们发现一方面两学期成绩情况与 Target 之间存在明显较强的相

关性（热图的最后一行）这表明学术表现是影响学生辍学与否的核心因素之一。而另一方面涉及两学期成绩的若干列数据维度中的属性之间的相关性系数也较高，这说明属性间的存在明显的多重共线性的影响，因此我们考虑将两学期以“Curricular_units_1st”与“Curricular_units_2nd”为前缀的数据合并取平均，以减轻数据集的多重共线性。

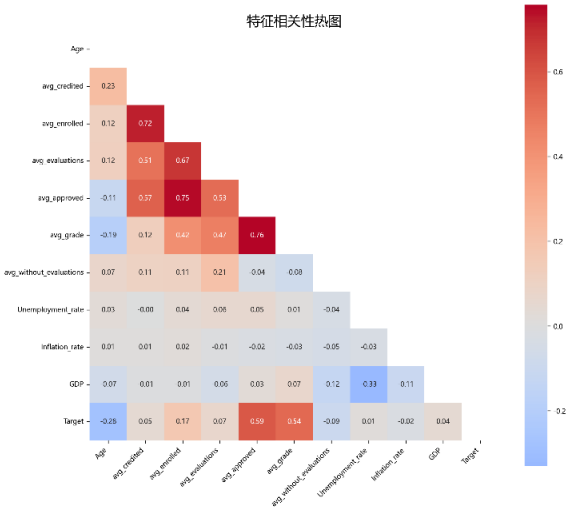

图 2 处理两学期成绩后的特征相关性热图

可以看出，处理后的多重共线性有所减轻，但对线性模型仍有较大影响，因此在后续的模型选择中我们需要考虑到这个问题。

上述分析过程提供了特征间最基础的线性关系理解，但由于其无法捕捉非线性关系，因此接下来我们将结合其他模型进行进一步验证。

五、 方法与算法

5.1 Baseline 方法

在本研究中，我们所采用的数据集为 kaggle 上开源的“Gold”数据集，voting 量高达 465，访问量 248k 次，下载量 41.5k 次。我们在其中挑选了一篇收获银牌的优质公开笔记本《ML Algorithms Usage and Prediction》（voting 量 43 次）作为基准模型（Baseline），以便与我们构建的模型进行性能对比。该笔记本由 Sunayana Gawde 创建，目标是评估多种机器学习算法在学生数据集上的表现。在笔记本中作者应用了包括 K 近邻、支持向量机和随机森林等常见分类模型，分别进行了三分类以及二分类的任务，并采用准确率、精确率、召回率和 F1 分数等指标对模型性能进行了比较，最终选择了调参后的随机森林作为最终模型。通过与该基准模型的对比，我们能够更客观地评估所提出模型的

优势和改进空间，从而为学生成绩预测提供更有效的解决方案。

5.2 模型方法设计

从之前的分析中，我们首先注意到数据集在类别分布上呈现显著不平衡特征，数据集中包含 4,424 条学生记录，其分布严重倾斜：研究生类别作为多数类占总体的 50%，辍学类别占 32%而入学类别仅占 18%，这对我们的分类任务提出了挑战。

为了解决数据不平衡的问题，我们引入了合成少数类过采样技术[9]（Synthetic Minority Over-sampling Technique, SMOTE）。SMOTE 是一种基于特征空间的过采样方法，其核心思想是在少数类样本之间的特征空间中生成合成样本，而非简单复制已有样本。这种方法的优势在于，它不仅增加了少数类的样本数量，还扩展了少数类在特征空间中的覆盖范围，使决策边界更加平滑，降低了过拟合风险。与简单复制相比，SMOTE 生成的样本具有更好的多样性，能够帮助模型更好地学习少数类的特征分布。

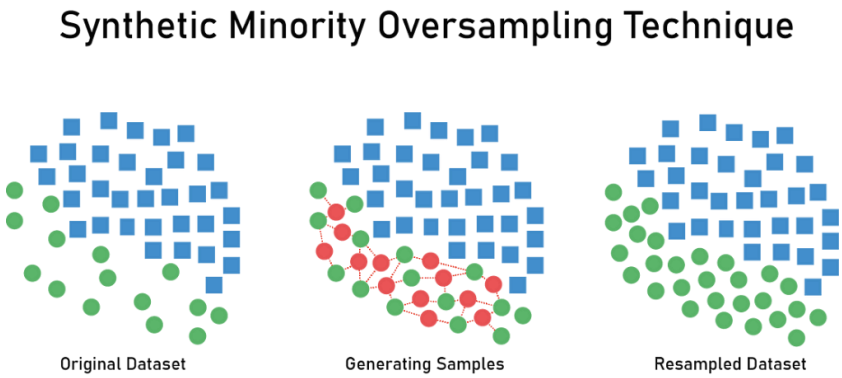

图 3 SMOTE 原理图

在模型设计与训练过程中，我们探索了四种不同的分类任务模型：(i) Logistic 回归线性模型 (ii) 随机森林集成模型 (iii) CatBoost 模型 (iv) LightGBM 模型,以期望从不同角度量化特征重要性，共同提供更全面可靠的分析结果，获取得到最佳的模型训练效果实践。我们首先采用逻辑回归模型,进行线性模型系数分析。该方法假设特征与目标变量间存在线性关系，首先帮助我们直观理解各特征对不同升学结果的影响方向与强度。我们接着构建了随机森林集成模型，进行随机森林特征重要性分析，以捕捉特征间的非线性交互作用。

在之前的分析中我们注意到属性间存在明显的多重共线性，并且许多原始属性的数值仅表示分类，这些特性对上述两种模型均构成了挑战：逻辑回归受多重共线性影响显著；而随机森林的特征重要性可能被高基数属性偏置。因此在接下来的分析中我们引入了两个能够处理属性间多重共线性且支持原始类别变量的模型：CatBoost[10]以及

LightGBM[11]。CatBoost 模型是一种专门优化的梯度提升决策树算法。该模型原生支持类别型特征处理，无需手动编码转换，最大程度保留原始信息结构。而 LightGBM 模型采用基于直方图的决策树学习与 Leaf-wise 生长策略，在保证模型精度的同时显著提升训练效率。这两个模型能为我们提供独特的特征重要性衡量标准。

为了进一步提升模型性能，我们同时对随机森林、CatBoost 以及 LightGBM 三种复杂模型采用了 RandomizedSearchCV 库进行超参数优化。最终我们训练得到了 7 个不同的模型：逻辑回归（LR）、随机森林（RF）、随机森林+参数微调(tuned_RF)、CatBoost（CB）、CatBoost+参数微调(tuned_CB)、LightGBM（LGBM）、LightGBM+参数微调(tuned_LGBM)。对每个模型，我们统一记录和分析以下性能指标： 准确率(Accuracy)、加权平均精确率(weighted precision)，加权平均 F1 分数(weighted f1 score)，加权平均召回率(weighted recall)。注意到 Enrolled 的分类样本数量少，分布不均衡，我们还额外评估了在 Enrolled 分类上的召回率(Enrolled recall)，在 Enrolled 分类上的 f1 分数(Enrolled f1-score)，从而全面评估各模型的预测效果和特征贡献。

5.3 三分类模型效果

我们所采用的 Baseline 和最终训练得到的 7 个模型的训练指标如下，用粗体标识表现最好的模型，用下划线标识表现第二好的模型：

表 1 三分类模型效果对比表

	Accuracy	Weighted precision	Weighted f1 score	Weighted recall	Enrolled recall	Enrolled f1-score
Baseline	<u>0.7816</u>	<u>0.7796</u>	**0.8014**	0.7630	0.4325	0.5169
LR	0.7612	0.7319	0.7333	0.7656	0.2289	0.3014
RF	0.7578	0.7340	0.7461	0.7524	0.3638	0.4323
tuned_RF	0.7713	0.7525	0.7544	0.7689	0.2844	0.3818
CB	0.7679	0.7742	0.7677	0.7679	**0.5390**	<u>0.5171</u>
tuned_CB	0.7762	0.7702	0.7690	0.7762	0.4610	0.4965
LGBM	0.7798	0.7775	0.7749	0.7798	0.4935	0.5135
tuned_LGBM (ours, best)	**0.7940**	**0.7882**	<u>0.7883</u>	**0.7940**	<u>0.5000</u>	**0.5329**

从整体表现来看，基于梯度提升的模型（CB、tuned_CB、LGBM、tuned_LGBM）明显优于传统模型。其中，tuned_LGBM 在加权准确率、精确率、F1 分数和召回率上均达到 0.79，展现出最为平衡和稳健的性能，为本研究中综合表现最佳的模型。相较之下，逻辑回归与随机森林在加权指标上的表现均弱于 Baseline，其中逻辑回归模型再“Enrolled”类的召回率和 F1 值方面表现最差,显示线性模型难以有效捕捉复杂关系。

值得注意的是，虽然 Baseline 模型在加权 F1 上略高，但在对“Enrolled”这一重要少数类的识别上表现有限（召回率为 0.43，F1 为 0.51），而 CB 模型在“Enrolled”召回率上达到 0.54，tuned_LGBM 在其 F1 分数上达到 0.53，均优于 Baseline，说明我们训练的模型在保持整体准确率的同时，更加关注少数类的识别能力，这与我们进行的 SMOTE 优化密不可分。

我们同时注意到模型调优带来不同程度的改善。例如，tuned_RF 在整体准确率上优于 RF（0.77 vs 0.75），但在“Enrolled”识别能力上出现下降，说明参数调优过程中存在向多数类倾斜的风险。相比之下，tuned_LGBM 在加权指标和少数类性能上均优于基础 LGBM，实现了性能全面提升。从模型结构角度分析，梯度提升类模型对类别变量与非线性特征处理更为出色，尤其是 CatBoost 在不需要独热编码的前提下能直接处理类别型变量，表现出较强的适应性与解释性。相比之下，传统模型受限于算法机制，在处理类别不平衡和多维变量时存在明显短板。

综上所述，我们通过 Baseline 与七种模型的系统比较，验证了集成学习方法（特别是 LightGBM）在学生升学状态预测任务中的优势，尤其在识别潜在中途辍学风险人群方面展现出更高的实用价值。

5.3 二分类模型效果

在三分类模型的训练结果中，我们注意到所有模型在对“Enrolled”（在读）类别的预测表现均不理想,该问题引发了我们对三分类建模方式本身的反思。从语义上看，“毕业（Graduated）”与“辍学（Dropout）”是学生最终的确定性升学状态，而“在读（Enrolled）”则属于尚未完成的中间状态，在模型训练过程中这种中间类别不仅容易被误判，还可能为模型引入噪声，进而影响整体分类性能。另一方面，从研究目标来看，我们更关心学生是否存在辍学风险，以便为潜在高风险群体提供支持。因此，将任务简化为“辍学”与“未辍学”（毕业）之间的二分类，不仅更符合实际干预需求，也可能带来更清晰、更稳定的模型表现。

基于此假设，我们构建了四个二分类模型：CatBoost（CB_bi）、LightGBM（LGBM_bi）以及它们的参数优化版本 tuned_CB_bi 与 tuned_LGBM_bi。在数据预处理阶段，我们移除了所有标签为“Enrolled”的样本，仅保留“Dropout”与“Graduated”两个类，并保持与三分类阶段一致的 SMOTE 过采样策略及评估指标体系。我们也设置了 Baseline_bi 模型用于横向对比，其训练结果来源于公开 Notebook 所提供的二分类模型

性能。

表 2 二分类模型效果对比表

	Accuracy	Weighted precision	Weighted f1 score	Weighted recall
Baseline_bi	0.8769	0.8852	0.8834	0.8962
CB_bi	0.9054	0.9054	0.9046	0.9054
tuned_CB_bi	**0.9185**	**0.9188**	**0.9178**	**0.9185**
LGBM_bi	0.9083	0.9098	0.9070	0.9083
tuned_LGBM_bi	0.9170	**0.9188**	**0.9178**	**0.9185**

最终结果如表所示，四个模型的准确率均超过 90%，远超三分类任务中最佳模型（LGBM）的 80%。其中，tuned_CB_bi 模型在准确率（0.9185）、加权精确率（0.9188）、F1 分数（0.9178）与召回率（0.9185）等指标上均排名第一，为整体性能最优。

综上，我们认为将问题建模为二分类任务，不仅在理论上更具可解释性，也在实践中显著提升了模型性能。同时，CatBoost 和 LightGBM 在保持高准确率的同时，也展现出较强的泛化能力和可移植性，为学生辍学风险建模提供了高效、可靠的解决方案。

六、 实验结果

从前述模型对比中可看出，我们训练的 tuned_CB_bi 在所有分类模型中效果最好，下面我们针对这一模型来分析影响辍学率的因素。首先我们进行实验结果的可视化，给出总体的影响因素柱状图。

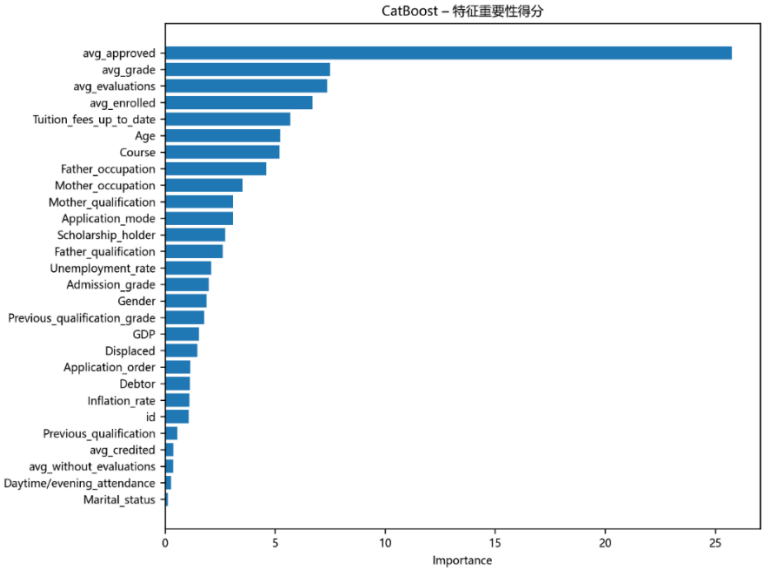

图 4 tuned_CB_bi 特征重要性划分柱状图

从图中可以看出，影响学生辍学率最重要的几个因素有：“avg_approved”“avg_grade”“avg_evaluations”“avg_enrolled”等成绩因素，随后是“Tuition_fees_up_to_date”“Age”“Course”等个人和家庭因素。下面我们来逐个分析其中的关系。

在此前的数据分析中，我们已经观察到“avg_approved”和“avg_grade”等变量与学生在课程结束时的情况（“辍学”、“在读”或“毕业”）之间的关联。现在，让我们仅针对“辍学”和“毕业”这两个标签，来研究三个重要的课程单元相关变量。

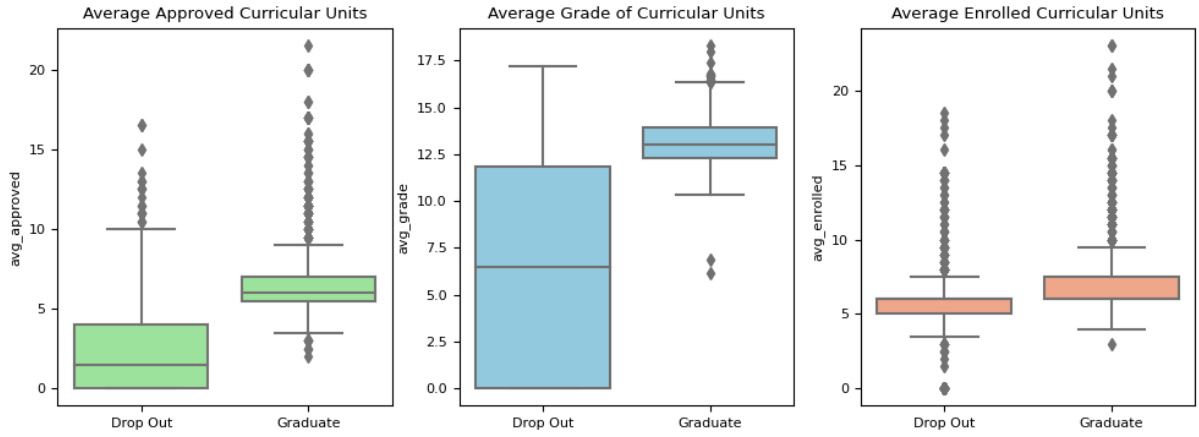

图 5 辍学与否和学生成绩、已修学分、注册学分的箱型图

可以看出，在正常学制结束时毕业的学生，其已经修满的课程学分数量更多，且第一和第二学期的平均成绩也远高于中途辍学的学生，说明众多学生因为学习成绩无法跟上而被学校劝退或主动退学。

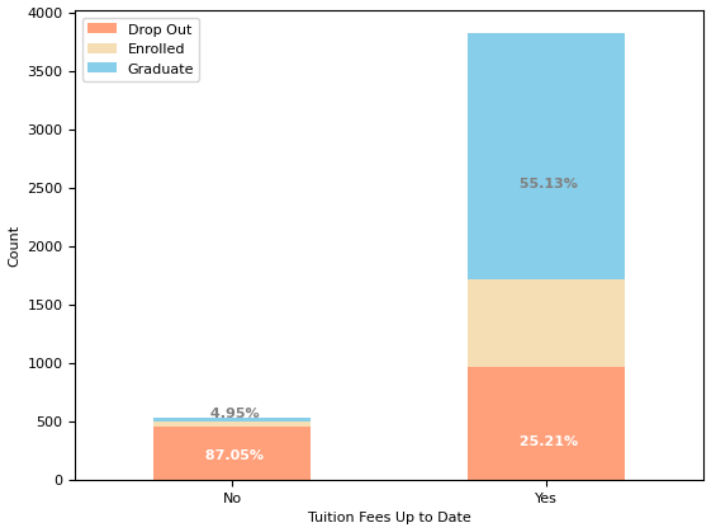

图 6 是否缴清学费与学生辍学关系的堆叠图

可以看出，那些学费没能按时缴纳的学生辍学率高达 87.05%，而毕业率则低至

4.95%，由于经济状况等原因无法按时缴清学费的确是辍学的一大主因。

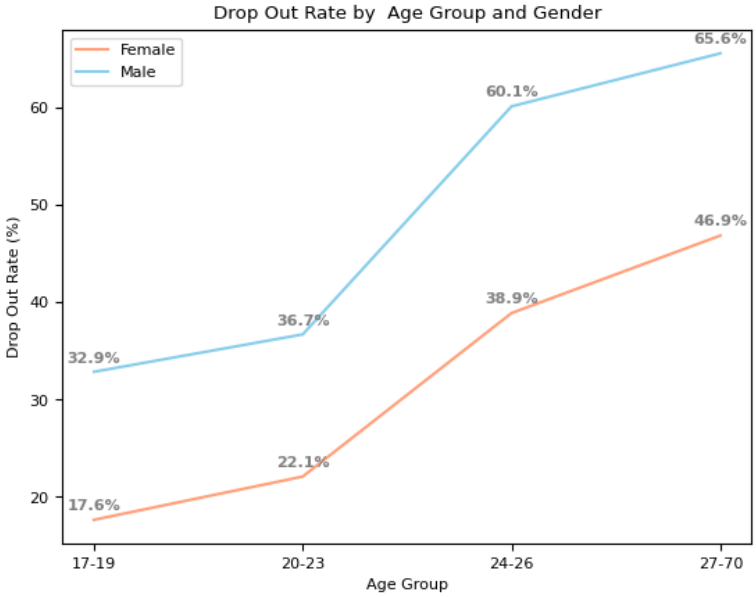

图 7 年龄、性别与辍学率关系折线图

从图中可以看出，无论是女性还是男性，其辍学率都会随着年龄的增长而上升，但总体而言，女性在所有年龄段的辍学率都低于男性，这表明年龄和性别都是影响学生辍学可能性的重要因素。

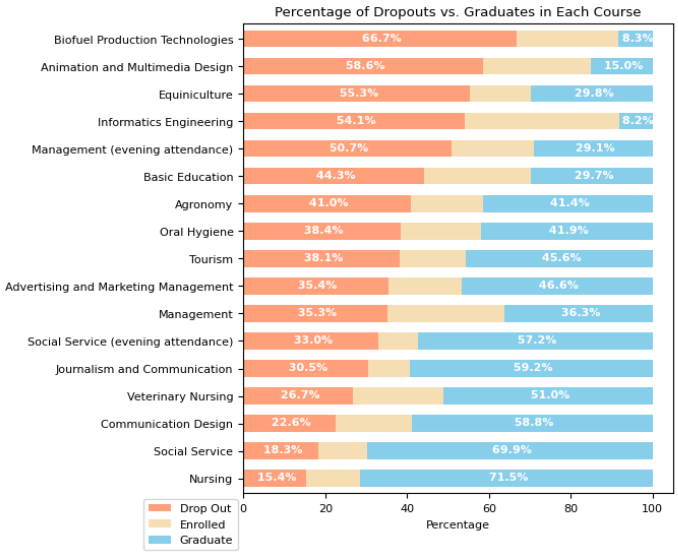

图 8 每门专业的辍学率百分比

不同专业的辍学率在 15.4% 至 66.7% 之间不等。在 17 门专业中，有 5 门的辍学率超过了 50%。生物燃料生产技术专业的辍学率最高，紧随其后的是多媒体设计、水产养殖专业、信息工程专业、管理学等；社会服务、护理专业的辍学率最低。大体上讲，专业学习难度越高，学生越容易辍学。

七、 结论

本研究围绕学生升学状态（辍学、在读、毕业）的预测与分析展开，系统地完成了从数据预处理、特征分析、模型训练到结果可视化的全过程。在此基础上，我们得出以下几个核心结论：

1. 影响学生升学状态的因素可大致划分为两类：课程相关表现和个体经济背景。其中，“avg_approved”、“avg_grade”、“avg_evaluations”、“avg_enrolled”等与学业成绩密切相关的变量占据主导地位，表明学生在校期间的学术表现直接决定其辍学风险。而“Tuition_fees_up_to_date”、“Age”、“Course”则紧随其后，凸显经济条件与学科难度对辍学倾向的影响。具体分析发现：未能按时缴清学费的学生辍学率高达 87.05%，远高于其他学生，说明经济压力是促使学生辍学的重要因素之一。与此同时，学生年龄越大，其辍学率越高，且男性辍学率在所有年龄段普遍高于女性，反映出性别与年龄的交互效应。我们还发现不同专业的辍学率差异也较为明显，其中生物燃料、多媒体设计、水产养殖、信息工程等专业辍学率显著偏高，提示部分课程内容难度与就业前景可能间接影响学生选择和坚持的意愿。

2. 在模型表现方面，我们认识到线性模型 Logistic 回归虽然具备良好的可解释性，但面对多重共线性与非线性特征时显得力不从心；而随机森林尽管在捕捉非线性关系上有所提升，但其特征重要性易受高基数类别属性影响。相比之下，CatBoost 与 LightGBM 模型在本任务中表现最优，准确率和 F1 均超过 0.77，且具有良好的处理类别型变量和不平衡数据的能力，适合复杂教育数据的建模任务。我们还发现，SMOTE 过采样技术在一定程度上缓解了样本类别不平衡问题，结合合理的参数调优策略可进一步提升模型性能。

3. 相较三分类任务，我们进一步明确：将问题转化为二分类任务（辍学 vs 未辍学）更贴近教育干预实践，也显著提升了模型性能。在二分类任务中，所有模型的准确率均突破 90%，远高于三分类任务的上限。tuned_CB_bi 模型在准确率、F1 分数和召回率方面均表现最优，成为本研究综合性能最强的模型。

4. 尽管本研究取得了初步成果，但仍存在若干不足之处。一方面，数据集样本量相对较小，类别不平衡问题突出，尽管通过 SMOTE 等方法进行调整，但仍可能影响模型的泛化能力与稳定性；另一方面，部分变量的编码较为粗略（如父母职业），由于我们缺乏葡萄牙本地的领域知识，无法合理分组或排序，因此，大多数多类别变量与

目标变量的分析跳过，未能充分挖掘其潜在层次结构。此外，本研究仍基于静态特征建模，未来可引入时间序列数据与学生行为轨迹，开展动态建模与长期跟踪分析，以获得更具前瞻性的预测能力。

基于上述研究发现与分析结果，我们向各利益相关方提出如下建议：

- **对学生个人**：应尽早关注自身的学习状态，特别是在课程成绩未达标或经济压力较大时，主动寻求学业和心理支持，避免小问题积累为辍学风险。
- **对家庭**：建议家庭成员在经济和情感上为学生提供持续的稳定支持，主动关注学生的心理状态与学习进展，营造良好的学习氛围。
- **对学校**：应建设基于数据驱动的学生预警与干预系统，结合课程不及格、经济困难、心理异常等多源信息，及时识别高风险学生群体并开展个性化帮扶。
- **对政府与社会**：应完善奖助学体系，加大对经济困难学生的资助力度，推动高等教育资源公平配置，缓解因社会结构性不平等引发的辍学现象。

通过多方协同努力，辅以科学的预测模型与决策机制，我们有望进一步降低辍学率、提升教育公平性，为高等教育可持续发展提供技术与数据支持。

参考文献

[1] Kehm, B.M.; Larsen, M.R.; Sommersel, H.B. Student Dropout from Universities in Europe: A Review of Empirical Literature. *Hungarian Educ. Res. J.* **2020**, *9*, 147–164.

[2] Atchley, W.; Wingenbach, G.; Akers, C. Comparison of Course Completion and Student Performance through Online and Traditional Courses. *Int. Rev. Res. Open Distance Learn.* **2013**, *14*, 104–116.

[3] Quinn, J. *Dropout and Completion in Higher Education in Europe among Students from Under-Represented Groups*; An Independent report authored for the NESET network of experts; European Commission: Brussels, Belgium, 2013.

[4] Namoun, A.; Alshanqiti, A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Appl. Sci.* **2020**, *11*, 237.

[5] Saa, A.A.; Al-Emran, M.; Shaalan, K. Mining Student Information System Records to Predict Students' Academic Performance. *Adv. Intell. Syst. Comput.* **2020**, *921*, 229–239.

[6] Akçapınar, G.; Altun, A.; As ̧kar, P. Using Learning Analytics to Develop Early-Warning System for at-Risk Students. *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 40.

[7] Daud, A.; Lytras, M.D.; Aljohani, N.R.; Abbas, F.; Abbasi, R.A.; Alowibdi, J.S. Predicting Student Performance Using Advanced Learning Analytics. In Proceedings of the 26th International World Wide Web Conference 2017, WWW 2017 Companion, Perth, Australia, 3–7 April 2017; pp. 415–421.

[8] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, *7*(11), 146. https://doi.org/10.3390/data7110146

[9] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

[10] Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *arXiv* **2017**, arXiv:1706.09516v5.

[11] Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018.