校园卡里的〝精打细算〞密码：多场景消费如何解码〝经济隐形人〞？

摘要

在社会经济发展背景下,高校仍有部分学生因家庭经济困难面临生活与教育支出的双重压力，当前高校贫困生资助工作也存在传统认定方式痛点。现如今，大数据技术在“精准资助”中大有潜力。本文借助某高校学生的校园卡消费数据，尝试探索构建可识别困难学生并保障隐私的隐性资助机制,具体开展以下三方面研究：

针对问题一，先对所给的原始数据进行预处理，将消费场所划分为“核心场所”和“辅助场所”，提取并可视化不同消费场所的消费特征。再按性别划分学生群体，借助频率分布直方图与第 5 百分位数识别异常消费行为，初步勾勒经济困难学生“低消高频、场景受限、金额紧缩”的画像。

针对问题二，以问题一的三个特征为一级指标，选取单次平均消费金额、消费频次等 7 个可量化特征作为二级指标。引入熵权-TOPSIS 综合评价模型，经熵权法赋权、TOPSIS 法计算贫困得分,排序后筛选出得分位于后 50 名学生的名单。

针对问题三，基于问题二，选用 K-means 聚类算法对这 50 名学生进行分类。由肘部法则进确定最优聚类数 K=4，将 50 名学生划分成 4 个资助等级。随后，估算学生的基本生活消费需求，以特困组学期消费估算为基准，设置其他等级的资助标准，形成梯度资助结构，最终测算出各级科学资助的金额区间分别为 2664.8～2764.8 元、2418.0～2518.0 元、2171.2～2271.2 元、1924.4～2024.4 元。

关键词：校园精准资助；熵值-TOPSIS 法；K-means 聚类

目 录

一、自拟题目描述

随着社会经济发展，人民生活水平整体提升，但在高校群体中，仍存在部分学生因家庭经济困难，其生活质量与教育支出承受能力显著低于校园平均水平。这些学生，特别是来自低收入家庭的大学生，常需在学业之外承受巨大的经济压力，为减轻家庭负担而不得不大幅压缩日常开支，其消费行为模式呈现出明显区别于普通学生的特征。当前高校贫困生资助工作面临诸多挑战：传统认定方式存在认定精度有限、易受主观因素影响、涉及隐私暴露、易引发心理负担等痛点；同时，资助资源有限性与需求广泛性之间的矛盾，亟需精准化、科学化、动态化的资助政策来提升资源配置效率与公平性[1]。在此背景下，大数据技术在促进教育公平、实现“精准资助”方面展现出巨大潜力[2]。我们将借助某高校300余名学生产生的逾2万条校园卡消费流水数据，运用大数据分析技术与数学建模方法，探索构建一种既能有效识别经济困难学生[3]，又能充分保障其个人尊严与隐私的“隐形资助”机制。基于上述背景，我们拟定研究以下问题：

（1）分析学生在不同消费场所的活跃用户规模、消费频次分布、人均消费水平等统计特征，并识别单次消费金额显著低于该场景常态水平的异常消费行为模式。

（2）在（1）的分析结果中建立模型，从研究样本中识别并筛选出经济困难程度最高的前50名学生作为最需要资助的候选对象。

（3）划分最需要资助的50名学生的资助金额等级，测算科学资助的金额区间。

二、问题分析

2.1 问题一分析

对于问题一，首先需要对现有数据进行整理，筛选出关键的数据，排除干扰，并按消费场所进行划分。由于消费场所较多，可将其分为“核心场所”和“辅助场所”两部分进行探究。在此基础上，借助可视化工具，呈现出不同消费场所的活跃用户规模、消费频次分布、人均消费水平的情况，以把握不同场所的消费水平特点。接着，借助频率分布直方图与百分位数，捕捉偏离数据主体分布的异常值，获取单次消费金额显著低于该场景常态水平的异常消费行为。最后，尝试对经济困难学生进行画像。

2.2 问题二分析

为精准识别最需要资助的学生，首先基于问题一概括的贫困生画像特征，选取相对应的可量化指标，建立评价指标体系。接着引入熵权-TOPSIS 综合评价模型，通过熵权法计算各二级指标的权重，再用 TOPSIS 法计算相对接近度作为贫困得分，分值越低表明经济困难程度越高，据此对学生的经济困难程度进行排序，筛选出得分位于后 50 名学生的名单。

2.3 问题三分析

根据问题二确定的 50 名学生名单，选取 K-means 聚类算法对这 50 名学生进行分类。通过肘部法则进行评估，确定最优聚类数，对该 50 名学生划分资助等级。随后，估算学生的基本生活消费需求，再以特困组的学期消费估算为基准，设置其他等级的资助标准，形成梯度资助结构，最终测算出科学资助的金额区间。

2.4 探究流程

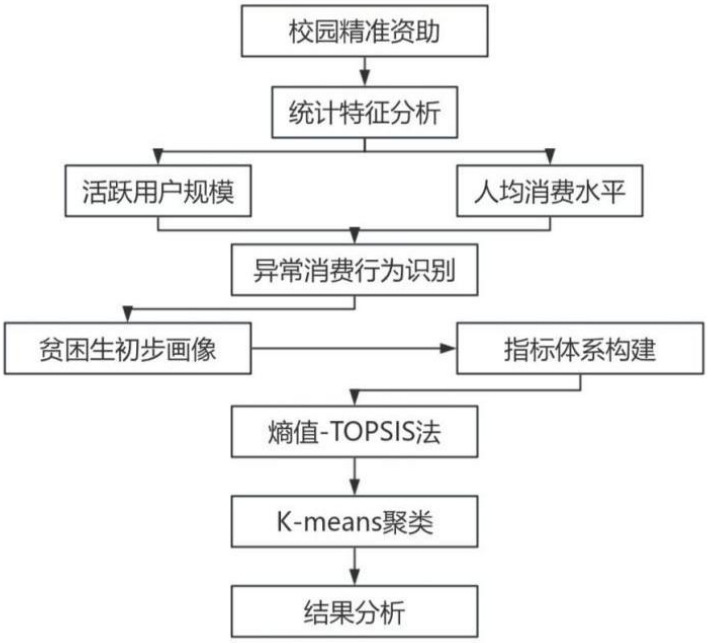

图 2-1 探究流程图

三、模型假设

1. 所有消费场所均 24 小时营业，各消费场所的常态消费水平在研究周期内相对稳定；

2. 每一位学生对应一张校园卡，默认一张卡仅由该学生本人使用；

3. 在给定数据所示的时间范围内，学生因家庭经济状况导致的消费行为模式较为稳定，不会出现大幅、持续性波动；

4. 学生在校园内的消费行为是理性决策结果，而非因消费习惯不良等非理性因素导致。

四、定义与符号说明

表 4-1 符号的定义与说明

符号定义	符号说明	符号定义	符号说明
C_i	第 i 个学生的消费总额	x_{ij}	第 i 个对象、第 j 个评价指标

n_i	第 i 个学生的消费次数	p_{ij}	第 i 个对象、第 j 个评价指标的比重
x_i	第 i 个学生的单次平均消费金额	H_j	第 j 项指标的熵值
X	单次平均消费金额数据集	w_j	第 j 个指标的熵权
R	极差	z_{ij}	第 i 个对象、第 j 个评价指标归一化
k	组数	Z^*	加权矩阵
h	组距	D	距离
A	初始矩阵	C_i	第 i 个对象相对接近度

五、模型的建立与求解

5.1 问题一的模型建立与求解

5.1.1 数据预处理

本文使用的数据为某高校329名学生产生的23637条校园卡消费记录表与学生ID表。其中，消费记录表中包含：消费时间、消费金额、存储金额、消费次数等十项信息,消费记录表中“消费项目的序列号”字段有23528条信息为NULL值，占该字段数据的99.58%，该字段对消费特征分析无实质影响，因此做剔除处理。

观察到在消费记录表中，消费类型共分成了存款、退款、无卡销户、消费四大类型。

表 5-1 消费类型

消费类型	存款	退款	无卡销户	消费
数据量（条）	747	6	7	22877

其中，存款为账户充值行为，本文主要研究的是学生的消费行为，该行为与消费场所的消费行为无关，若混入消费数据会导致统计偏差，因此将该数据单独存放，后续或可用于辅助验证。退款为消费后的金额返还，属于消费场景的衍生

行为，需保留分析，将退款数据与前一次消费数据做对比，需消去相对应的退款与消费情况。无卡销户为账户终止行为，与日常消费行为无关，因此做剔除处理。由于学生 ID 表中包含学生的性别信息，不同性别在消费习惯上可能存在固有差异，因此使用 VLOOKUP 函数将性别信息整合到消费记录表中。将消费记录表的数据按照“校园卡号-日期-时间”的顺序重新排列。观察发现，“消费次数”字段与给出的具体时间顺序存在不匹配的情况，且该消费次数应为总消费次数，而非记录表中给出的四月份消费累计次数，因此对该字段数据单独存放，后续或可用于辅助验证。同时，还发现部分学生消费数据显示在同一地点和同一时间段（1 小时以内）进行了多次消费，可能是由特殊情况（如多窗口消费、分次购买商品等）所导致的。因此，对这部分消费数据需进行合并处理。最终得到了 322 名学生产生的共 13854 条有效数据。

5.1.2 统计特征分析

(一) 不同消费场所下活跃用户规模

由于学生与校园卡号是一一对应的，所以我们可以根据整理好的数据，以四月份整个月为统计区间，通过识别各个场所中校园卡号出现的总个数，以此反映各个消费场所活跃的用户数量，得到图 5-1。

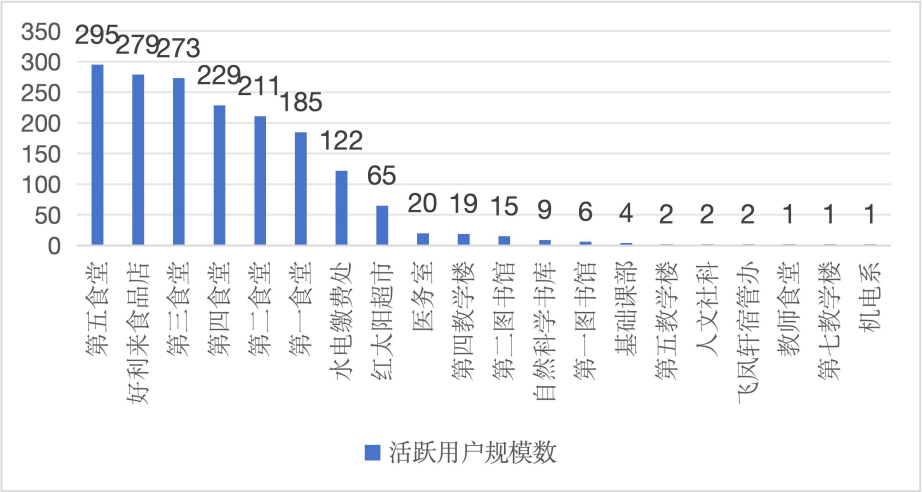

图 5-1 不同消费场所下活跃用户规模

从图中可看出，第五食堂、好利来食品店、第三食堂、第四食堂、第二食堂和第一食堂的活跃用户规模人数均大于 180 人，每个场所用户量占总人数的 56.23%以上，规模较大。并且，这 6 个消费场所均为餐饮服务场景，表明学生的日常生活核心需求主要集中在饮食方面，因此将这 6 个场所标记为“核心场所”。在核心场所中，虽然同为餐饮消费场所，但第五食堂、好利来食品店和第三食堂活跃的用户量明显高于第四食堂、第二食堂和第一食堂。好利来食品店作为非食堂场所，其用户量异常突出，表面学生零食消费需求强。

剩余的 14 个消费场所则标记为“辅助场所”，涵盖生活服务、医疗、教学等多种功能场地,与餐饮消费关联较弱,大部分场所的活跃用户量占比不到 20%，活跃用户规模普遍远低于核心场所。

（二）不同消费场所下消费频次分布

四月份各消费场所消费数据量呈现如下：

表 5-2 各消费场所消费数据量

序号	场所分类	地点	总数据量	占比	平均消费频次（次/月）
1	核心场所	第五食堂	5756	41.25%	17.88
2		第三食堂	2531	18.14%	7.86
3		好利来食品店	1915	13.72%	5.64
4		第二食堂	1400	10.03%	4.35
5		第四食堂	1211	8.68%	3.76
6		第一食堂	720	5.16%	2.24
7	辅助场所	水电缴费处	184	1.32%	0.57
8		红太阳超市	147	1.05%	0.46
9		医务室	23	0.16%	0.07
10		第四教学楼	21	0.15%	0.07
11		第二图书馆	16	0.11%	0.05
12		自然科学书库	9	0.06%	0.03
13		第一图书馆	6	0.04%	0.02
14		基础课部	4	0.03%	0.01
15		教师食堂	4	0.03%	0.01

16	第五教学楼	2	0.01%	0.01
17	人文社科	2	0.01%	0.01
18	第七教学楼	1	0.01%	0.00
19	飞凤轩宿管办	1	0.01%	0.00
20	机电系	1	0.01%	0.00

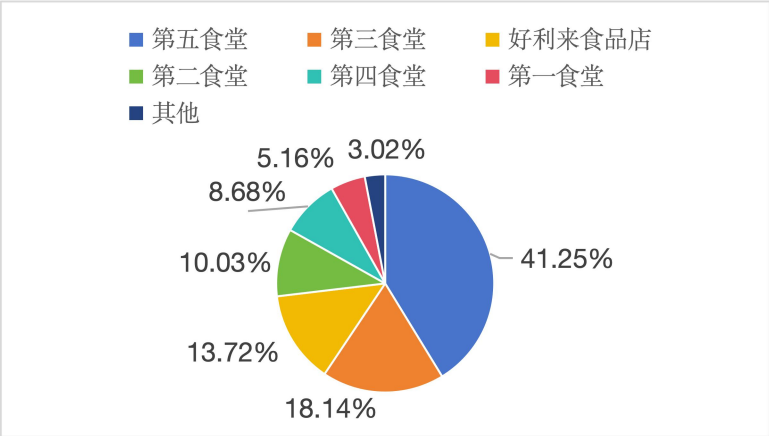

图 5-2 各消费场所消费数据占比

从图表中数据可以发现，核心场所数据总量占总体的 96.98%，是消费行为的集中发生地；辅助场所消费场景多为低频消费、功能性服务场景，其数据总量仅占总体的 3.02%，发生频率远低于餐饮场所。对于平均消费频次统计，其中核心场所消费频次合计 41.72 次/月，辅助场所消费频次合计 1.31 次/月，总体消费频次合计 43.03 次/月。

(三) 不同消费场所下人均消费水平

四月份各消费场所人均消费水平呈现如下：

表 5-3 各消费场所人均消费水平

地点	活跃用户规模数（人）	总计消费金额（元）	人均消费金额（元/人）
第五食堂	295	28356.41	96.12
好利来食品店	279	8903.20	31.91
第三食堂	273	19227.71	70.43
第四食堂	229	10518.00	45.93
第二食堂	211	7118.43	33.74
第一食堂	185	3833.40	20.72
水电缴费处	122	169.06	1.39
红太阳超市	65	732.00	11.26

医务室	20	182.10	9.11
第四教学楼	19	2290.00	120.53
第二图书馆	15	14.60	0.97
自然科学书库	9	8.70	0.97
第一图书馆	6	2.00	0.33
基础课部	4	21.00	5.25
第五教学楼	2	350.00	175.00
人文社科	2	9.00	4.50
飞凤轩宿管办	2	100.00	50.00
教师食堂	1	49.50	49.50
第七教学楼	1	50.00	50.00
机电系	1	300.00	300.00

从数据看，第五食堂无论是活跃用户数量还是总消费金额都位居榜首。在核心场所中第五食堂与第三食堂的人均消费金额处在较高水平。

5.1.3 异常消费行为识别

从样本构成上可以发现，在 322 个样本中，男生 44 人，女生 278 人，女生人数远远多于男生人数，反映该统计范围内女生消费参与群体规模更大。分别计算出男、女生两个群体的总消费金额与消费频次，得到男、女生平均消费频次与单次平均消费金额两组数据。数据显示，男生消费更倾向于单次较高金额支出，女生则为低单次、高消费频次的模式。

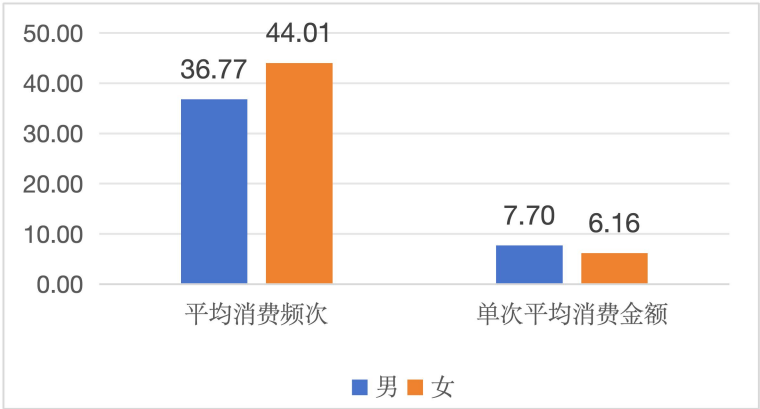

图 5-3 男、女生平均消费频次与单次平均消费金额

由于男、女生在消费模式中存在着显著差异，因此在识别异常消费行为时，

我们将学生群体按照性别进行区分，以规避群体特征混淆引发的判断偏误。频率分布直方图可以帮助我们直观地呈现数据分布形态，接下来我们将结合性别分组，分别绘制频率分布直方图。

（一）计算单次平均消费金额

设某性别群体中，第 i 个学生的消费总额为 C_i，消费次数为 n_i，计算出单次平均消费金额 x_i：

$$x_i = \frac{C_i}{n_i}$$

得到男、女生组单次平均消费金额数据集

$$X_{male} = \{x_1, x_2, \cdots, x_m\}(m = 44)$$

$$X_{female} = \{x_1, x_2, \cdots, x_n\}(n = 278)$$

（二）数据排序

对男生组、女生组的单次平均消费金额数据分别按从小到大的顺序排序，得到有序数据集 $X_{male,sorted}$ 和 $X_{female,sorted}$.通过对女生组消费数据分布的观察可知，其单次平均消费金额大多集中于 15 元以内，但存在 1 例异常值（单次平均消费金额为 50.1 ）。经追溯初始数据，该异常值对应的消费行为仅涉及第四教学楼与水电缴费处这两个辅助场所，与本研究聚焦的核心消费场景无关联，且其消费金额显著偏高，超出后续贫困生讨论范畴，故将该条数据予以剔除。

表 5-4 180158 号学生消费情况

校园卡号	性别	日期	时间	消费金额（元）	消费地点
180158	女	4 月 16 日	16:36	100	第四教学楼
180158	女	4 月 23 日	8:59	0.2	水电缴费处

（三）计算极差

极差反映数据的离散范围，公式为：

$$R = \max(X) - \min(X)$$

其中，$\max(X)$为数据集中的最大值，$\min(X)$为最小值，分别计算男生组极差 $R_{male} \approx 13.32$，女生组极差 $R_{female} \approx 11.89$.

（四）确定组数

Sturges 经验公式是统计学家斯特奇斯提出的，用于解决数据分组时组数确定的问题，可以依据样本量确定理论组数，为绘制频率分布直方图提供确定组数的参考标准。其计算公式为：

$$k = 1 + \log_2 N$$

其中，k 为组数，N 为样本量，分别计算可得男生组组数 $k_{male} \approx 7$，女生组组数 $k_{female} \approx 9$.

（五）计算组距

组距 h 是每组数据的区间长度，公式为

$$h = \frac{R}{k}$$

分别代入数据可得，男生组组距 $h_{male} \approx 2$，女生组组距 $h_{female} \approx 1.5$.

（六）划分区间、绘制频率分布直方图

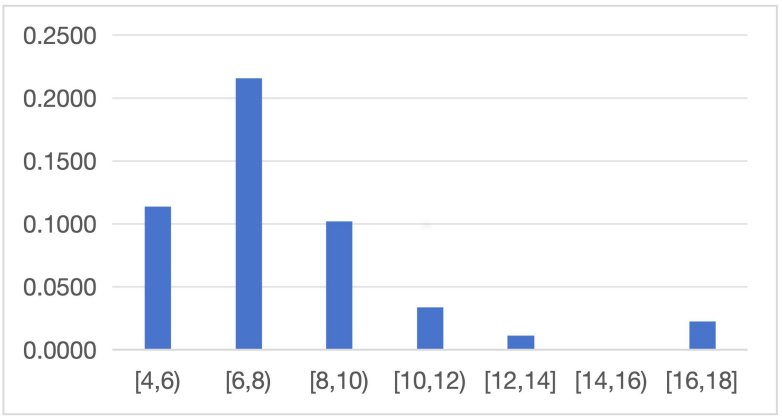

图 5-4 男生组频率分布直方图

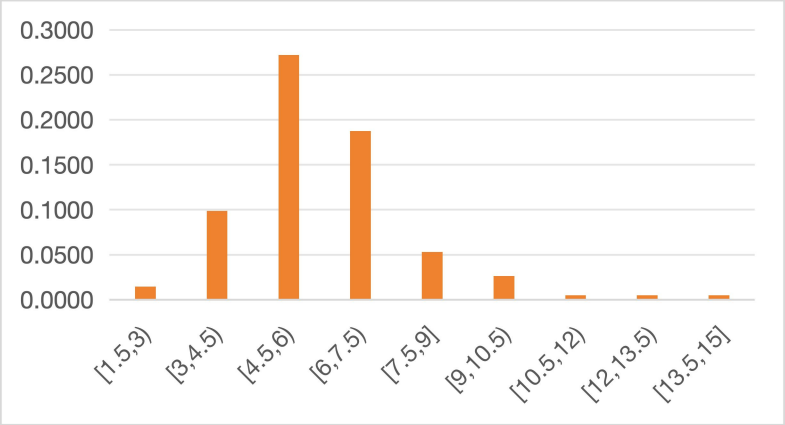

图 5-5 女生组频率分布直方图

由于四分位数可以将排序后的数据等分为四部分，其中，下四分位数代表处于 25%位置的数值，即有 75%的数据大于或等于这个数值，并且有 25%的数据小于或等于等这个数值，能够客观反映数据集中偏下的分布特征，可以帮助识别低于常态水平的具体情况。同时，对于异常低值的判断，在统计学中，通常会使用 5%分位数作为判断阈值，可用来描述显著低于该场景常态水平的异常消费行为模式。结合频率分布直方图，可得以下数据：

表 5-5 上四分位数与第 5 百分位数

	上四分位数	第 5 百分位数
男生组	6.11	4.44
女生组	4.80	3.29

观测核心场所异常低消情况发现，有 23 名学生在核心场所异常低消比例超过 50%，如下表所示。

表 5-6 核心场所异常低消比例超 50%情况

校园卡号	性别	核心场所异常低消比例	校园卡号	性别	核心场所异常低消比例
180212	男	62.50%	180242	女	55.56%
180225	男	57.69%	180118	女	54.72%
180340	男	55.56%	180159	女	54.69%
180136	女	82.69%	180111	女	53.33%
180061	女	81.82%	180137	女	53.13%

180382	女	74.14%	180089	女	52.78%
180192	女	69.14%	180252	女	52.63%
180360	女	69.09%	180359	女	52.63%
180381	女	62.26%	180141	女	52.38%
180063	女	61.29%	180178	女	52.00%
180387	女	58.33%	180117	女	51.85%
180184	女	57.14%			

从表中数据可以看出，仅有 7.14%学生异常低消占比超过 50%，女生数量相对较多，多数学生低消占比小于 50%。并且，可以发现异常低消比例数值存在一定跨度，最高异常低消比例为 180061 号女生，高达 81.69%，也存在相当一部分学生（17 人）最高异常低消比例低于 5%的情况。这表明，大部分学生消费行为符合常态水平，异常低消并非普遍现象，对后续深入挖掘需要给予资助的贫困生给予了可行的数据支撑。

5.1.4 贫困生初步画像

由于上述 23 名学生的异常低消占比较高，体现学生在核心消费场所频繁压缩单次消费，与贫困生因经济约束、主动选择低价消费以控制支出的行为逻辑高度契合，因此我们基于这 23 名学生校园消费数据，对贫困生进行初步画像。

我们具体提取了这 23 名学生的性别、总消费金额、日均消费、消费频次、单次平均消费金额、场所多样性等 10 项数据进行分析。下表仅展示部分数据详情：

表 5-7 异常低消超 50%的部分学生消费数据

序号	校园卡号	性别	总消费金额	日均消费	消费频次	单次平均消费金额
1	180136	女	126.90	4.23	52	2.44
2	180061	女	223.10	7.44	81	2.75
3	180212	男	112.90	3.76	24	4.70
4	180225	男	260.16	8.67	58	4.49

序号	校园卡号	场所多样性	核心场所低消频次	核心场所异常低消比例	核心场所消费频次	核心场所消费比

1	180136	5	43	82.69%	52	100.00%
2	180061	7	63	81.82%	77	95.06%
3	180212	4	15	62.50%	24	100.00%
4	180225	4	30	57.69%	52	89.66%

在消费金额与频次维度中发现，贫困生总消费金额与日均消费普遍处于较低水平，如校园卡号 180184 的学生，总消费仅 50.60 、日均消费 1.69 ，消费频次也相对偏少，反映出整体消费能力受限，需严格管控日常支出。在单次消费与场所多样方面，贫困生单次平均消费金额多集中在几元区间，像卡号 180136 的女生单次平均消费 2.44，且消费场景单一，卡号 180063 的女生场所多样值为 1，体现出因经济约束，倾向选择低价消费内容，消费选择范围受限。在异常低消特征上，贫困生异常低消频与比例显著偏高，卡号 180136 的女生异常低消频 43、比例 82.69%，核心场所低消频高且异常低消比例也高，说明即便在食堂等日常必需场景，也因经济压力持续低消，依赖低价选项维持基本生活。

综上，可以提取出贫困生的初步画像特征为：低消高频、场景受限、金额紧缩。

5.2 问题二的模型建立与求解

为实现对学生贫困程度的量化评估与分层资助识别，本文引入了熵权-TOPSIS 综合评价模型[4]，结合学生的消费行为特征对其经济困难程度进行评分排序。熵权法用于客观确定各指标的权重，通过计算各指标的离散程度，确定其在评分中的权重，避免人为赋值带来的主观偏差；TOPSIS 法则根据每位学生与“理想贫困状态”的相对距离进行排序，在标准化数据基础上，计算每位学生与最优、最劣解的欧氏距离，构造相对接近度得分，得出最终贫困分数，分值越低表示贫困程度越高。具体流程如下。

5.2.1 特征指标构建

在问题一中，我们得到了贫困生的初步画像。将概括出的三个特征作为一级指标，从消费行为的“金融、频次、场景分布”等维度，选用可量化的二级指标对应一级指标的特征，以此将抽象的概念转化为具体、可比的数据，具体指标选取呈现如下：

表 5-8 校园精准资助指标体系

一级指标	二级指标	单位	属性
低消高频	单次平均消费金额	元	反向
	消费频次	次	正向
	异常低消比例		正向
场景受限	场所多样性	个	反向
	必要消费占比（核心场所消费占比）		正向
金额紧缩	日均消费	元	反向
	消费金额波动率（CV）		反向

其中，消费金额波动率计算的核心是衡量消费金额在一定时间内的波动程度，在概率统计中，标准差与变异系数（CV）都可以用来刻画波动情况，但标准差只看绝对波动金额，会忽略学生本身的消费能力。而变异系数引入了均值，可以消除消费基线的差异，直接对比消费波动占其自身消费情况的比例。贫困生的消费种类和价格变化程度应比非贫困生小或者可选择消费的余地比较小[5]，因此消费金额波动率属性应设置为反向，即波动越小，贫困程度越高。

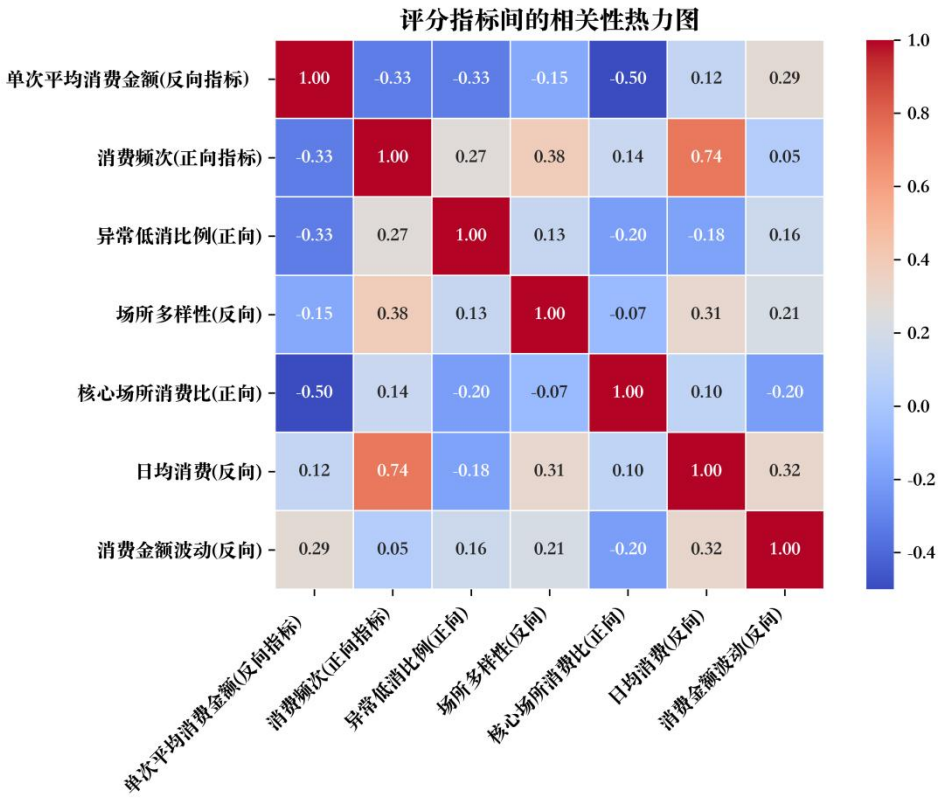

图 5-6 评分指标间的相关性热力图

为验证所构建特征指标的有效性，我们计算了各变量之间的皮尔逊相关系数，并绘制了特征指标间相关性热力图。从数据看，特征间关联呈现多元复杂态势，不同指标间相关系数有正有负、强弱各异，存在多样关联逻辑，具备一定的区分度，且部分指标相关性弱，体现出非冗余潜力。

具体来看，消费频次与日均消费（0.74）高度正相关，二者可以共同刻画学生的消费活跃度；单次平均消费金额与消费金额波动（0.29）有一定正相关，说明学生单次平均消费金额低且波动小，符合贫困生的消费相对稳定、金额不高的特点，可借此区分学生的消费稳定性；异常低消比例与消费频次（0.27）、场所多样性（0.13）等呈现弱正相关，说明三者之间存在一定关联特征，能补充识别消费行为的信息；核心场所消费比与单次平均消费金额（-0.50）的负相关明显，说明当学生核心场所消费比越高，单次平均消费金额就越低，符合贫困生的消费

特点。

综上,从指标的关联特征来看,这些指标能很好地为精准资助判定提供依据。

5.2.2 综合评价模型

在校园精准资助指标体系中，假设评价模型中有 n 个对象、m 个评价指标，建立初始矩阵 A。

$$A=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

（一）熵值法计算权重

1.将 x_{ij} 转化为比重形式 p_{ij}：

$$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}},\ i=1,2,\cdots,n;j=1,2,\cdots,m$$

2.计算第 j 项指标的熵值 H_j：

$$H_j=-\frac{\sum_{i=1}^{n} p_{ij}\ln p_{ij}}{\ln n},\ j=1,2,\cdots m$$

3.定义第 j 个指标的熵权为 w_j：

$$w_j=\frac{1-H_j}{n-\sum_{j=1}^{m} H_j},\ j=1,2,\cdots m$$

其中，$w_j\in[0,1]$，且 $\sum_{j=1}^{m} w_j=1$。

经过计算，得到赋权结果如下：

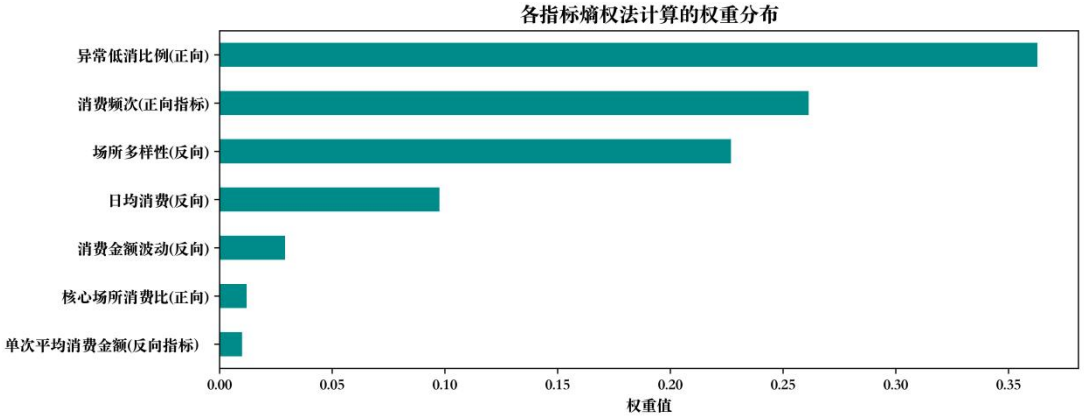

图 5-7 指标赋权

（二）TOPSIS 法计算综合得分指数

1.归一化

$$z_{ij}=\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}x_{ij}^{2}}},\ i=1,2,\cdots,n;\ j=1,2,\cdots,m$$

2.构造加权矩阵

$$z_{ij}^{*}=z_{ij}\times w_{j}$$

得到加权矩阵

$$Z^{*}=\begin{pmatrix} z_{11}w_{1} & z_{12}w_{2} & \cdots & z_{1m}w_{m} \\ z_{21}w_{1} & z_{22}w_{2} & \cdots & z_{2m}w_{m} \\ \vdots & \vdots & \cdots & \vdots \\ z_{n1}w_{1} & z_{n2}w_{2} & \cdots & z_{nm}w_{m} \end{pmatrix}$$

3.寻找最优、最劣方案

$$\begin{cases} z_{ij}^{*+}=\max\limits_{n,m}\left(z_{1}^{*+},z_{2}^{*+},\cdots,z_{m}^{*+}\right) \\ z_{ij}^{*-}=\min\limits_{n,m}\left(z_{1}^{*-},z_{2}^{*-},\cdots,z_{m}^{*-}\right) \end{cases}$$

4.最优、最劣距离

$$D_i^+ = \sqrt{\sum_j \left(z_{ij}^* - z_j^{*+}\right)^2}$$

$$D_i^- = \sqrt{\sum_j \left(z_{ij}^* - z_j^{*-}\right)^2}$$

5.构造相对接近度

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-},\ i = 1,2,\cdots,n$$

其中 $C_i \in [0,1]$，C_i 越大，综合得分越大，排名越高。

为进一步观察熵权-TOPSIS 模型在学生贫困识别中的区分效果，我们绘制了全体学生的综合得分分布直方图。

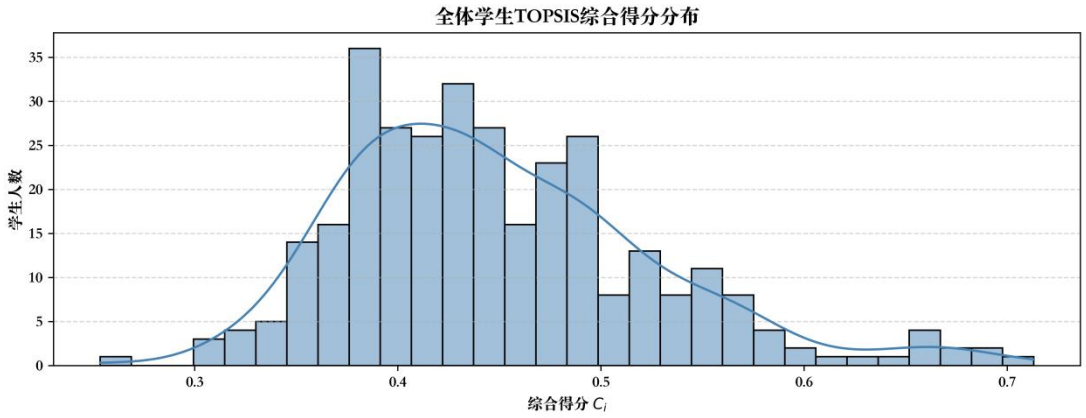

图 5-8 全体学生综合贫困得分分布图

该图象横轴为贫困综合得分，纵坐标为学生人数。得分越低对应贫困程度越高，得分越高对应经济条件越好。从图中分布我们可以发现，数据集中在[0.35,0.5]区间的频数较高，表明大量学生处于中等经济水平；[0,0.35]与[0.6,1]区间内的学生人数显著减少，表明经济条件极优或极差的学生占比较低。

该模型数据分布基本连续，输出稳定，整体曲线近似正态分布。故表明该模型能为筛选 50 名贫困生提供可靠依据。最终确定了 12 名男生、38 名女生为最需要资助的候选对象，具体名单如下：

表 5-9 最需要资助的 50 名学生名单

校园卡号	性别	校园卡号	性别	校园卡号	性别	校园卡号	性别	校园卡号	性别	校园卡号
180231	男	180293	女	180218	男	180074	男	180345	男	180231
180143	女	180053	女	180154	女	180112	女	180210	男	180143
180288	女	180295	女	180343	男	180146	女	180328	女	180288
180314	女	180363	女	180244	女	180336	女	180193	女	180314
180375	女	180095	女	180014	女	180162	女	180279	女	180375
180004	男	180309	女	180191	女	180264	女	180177	女	180004
180091	女	180036	女	180080	男	180166	女	180331	女	180091
180325	女	180099	女	180077	男	180236	男	180144	女	180325
180007	男	180358	女	180282	女	180150	女	180259	女	180007
180070	女	180045	女	180084	女	180226	男	180313	女	180070

5.3 问题三的模型建立与求解

5.3.1 K-means 聚类

为进一步挖掘这 50 名学生群体内部的贫困程度差异，本文采用 K-Means 聚类分析，对其消费特征进行无监督学习建模，以划分不同层级的资助等级。为确定最优聚类数 K，引入肘部法则（SSE），如下图所示，K=4 附近下降斜率开始变缓，肘部拐点明显。

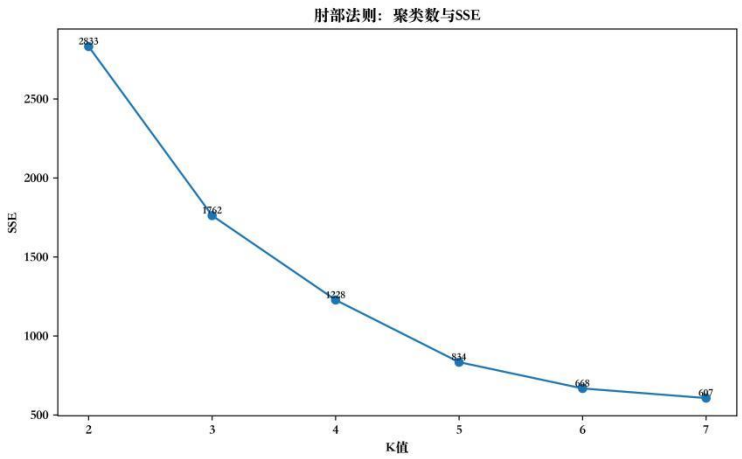

图 5-9 肘部法则

接下来对 50 名最需要资助的学生进行聚类，形成四类资助等级，其中 1 级

为特困等级，1-4 类贫困等级依次递减。该方法不依赖人工标签，完全基于学生的消费行为特征进行分类，具备良好的客观性和适应性。

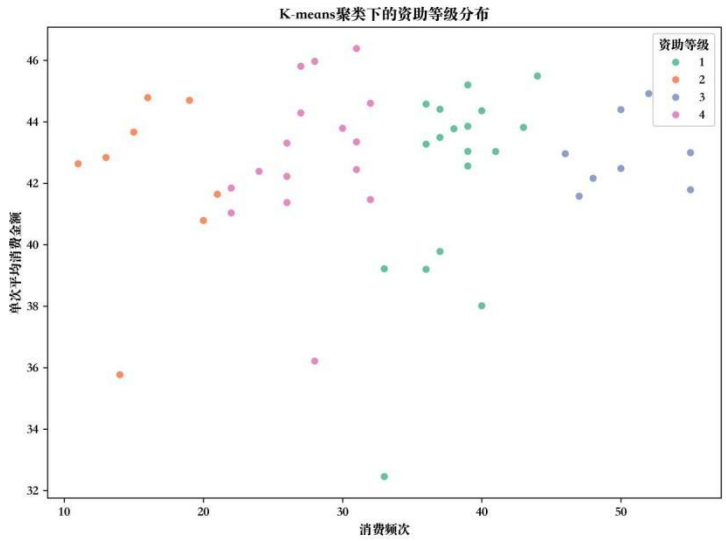

图 5-10 资助等级分布

从图中可以看出，1 级多分布在消费频次较高、单次平均消费金额偏低区域，体现出特困学生因经济压力，消费行为高频且小额的行为特点；2 级集中在中间区域，消费行为相对均衡；3、4 级的部分点有趋向消费频次低但单次金额偏高区域延伸，表明经济压力相对更小。

5.3.2 金额区间测算

在完成对资助对象的分级聚类后，为确保不同等级学生在经济上获得合理、差异化的资助支持，本文提出基于“消费能力反推保障线+分档补齐”相结合的测算方法，科学设定各等级学生的建议资助金额区间。具体做法如下：

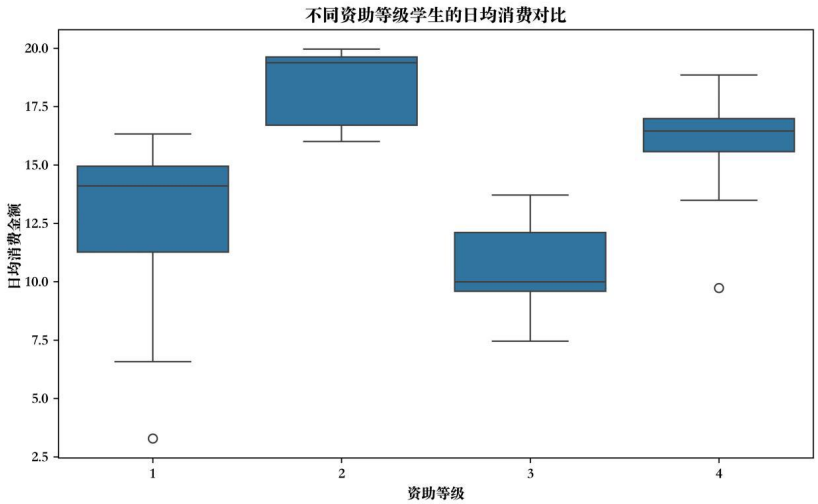

图 5-11 不同资助等级学生的日均消费对比

各资助等级学生的日均消费的具体分布如图，由于中位数相较均值更能代表该群体的典型水平，能避免因极端值干扰而失真。因此，我们以各等级群体的日均消费中位数×学期天数（150 天）作为对其基本生活消费需求的估算。随后，以最低等级（1 级）的学期消费估算为基准，在其基础上设置其他等级的资助标准，形成“上限略补、逐级递减”的梯度资助结构。

其中，1 级群体的日均消费中位数为 16.45 元，推算学期最低保障线为 2468 元，在此基础上给予 10%补贴，形成 2714.8 元建议资助金额，并设置“±50 元浮动”形成建议资助区间（2664.8~2764.8 元）。2 至 4 级群体分别给予基准金额（2468 元）的 100%、90%、80%的金额作为补贴，从而兼顾“保底需求”与“激励调节”。最终得到测算结果如表所示：

表 5-10 建议资助区间测算

资助等级	人数	日均消费中位数	估算学期消费	建议资助金额	建议资助区间
1	14	16.45	2468	2714.8	2664.8～2764.8 元
2	8	10	1499.25	2468	2418.0～2518.0 元
3	10	18.02	2703.5	2221.2	2171.2～2271.2 元
4	18	14.1	2114.5	1974.4	1924.4～2024.4 元

六、模型的评价及优化

6.1 模型的优点

传统认定方法常因依赖单一证明材料、认定方法简单粗放，难以全面反映学生真正经济状况，且缺乏动态监管机制，易出现“一认定终身”的僵化问题[6]。本文所建立的模型能通过消费数据的深度挖掘，将定性问题转化为定量问题，将认定过程从主观判断转向数据驱动的客观评估，突破了传统模式的局限性，为高校学生资助工作提供了更精准、灵活的解决方案。

6.2 模型的缺点

本文模型也存在着许多不足之处。首先，所依据的数据较老，可能与当前学生消费情况存在偏差。其次，在数据处理中，消费场所核心与辅助的划分、数据合并区间的选取等都依赖主观判断，缺乏统一客观的标准，可能影响对资助对象识别和等级划分的公正性。再者，本文所选数据维度较为单一，仅依赖学生在校的消费数据，缺乏学业表现、奖学金获取情况、勤工俭学参与记录、家庭背景信息等多维度数据的支撑。

6.3 模型的推广

该模型在高校学生资助工作中具有一定的应用推广价值，为精准资助提供了可操作的流程和方法。本文现有模型可加强结合学生的家庭背景、学习资源获取等情况，分析教育资源分配对不同群体学生发展的影响，从而能为教育政策调整提供一定的依据。本文的探究模型的核心逻辑是通过多维度数据的整合分析，实现对象特征的精准刻画、分类排序及差异化策略制定，这一思路亦可迁移到在校管理、公共服务、社会科学等多个领域的探究中。

参考文献

[1] 伍智鑫.基于一卡通数据挖掘的高校贫困生精准扶贫应用研究[D].安徽农业大学,2019.DOI:10.26919.

[2] 潘超,郭禹宏,穆宏浪.大数据下基于学生行为画像分析的高校精准资助模式构建研究[J].信息系统工程,2020,(12):74-76.

[3] 陈建顺,李照刚.高校贫困生的界定及其资助[J].曲靖师范学院学报,2003,(04):94-97.

[4] 孙群,郑丹妮.社会保障高质量发展的时空分异与动态演化研究——基于熵权 TOPSIS 法的测度分析[J/OL].重庆三峡学院学报,1-16[2025-07-24].

[5] 齐怀峰.大数据背景下高校贫困生类别的判定——以安徽师范大学为例[J].高校辅导员学刊,2016,8(05):74-77.DOI:10.13585.

[6] 程茜宇.基于深度神经网络的高校贫困生精准识别研究[D].江西财经大学,2019.

附　录

附录 1

问题二：Python 代码

特征相关性热力图

```
    corr = features.select_dtypes(include=["float64", "int64"]).corr()
    plt.figure(figsize=(12, 10))
    sns.heatmap(corr, annot=True, fmt=".2f", cmap="coolwarm")
    plt.xticks(rotation=45, ha='right')
    plt.title("特征相关性热力图")
    plt.tight_layout()
    plt.savefig("图 2_特征相关性热力图.png", dpi=600)
    plt.show()
```

各指标的熵权法权重分布

```
# === Part 4: 熵权法函数与正向化函数
     def transform_indicators(df, indicators, directions):
         df_new = df.copy()
      for col, direction in zip(indicators, directions):
        if direction == '负向':
          max_val = df[col].max()
          min_val = df[col].min()
       df_new[col] = max_val + min_val - df[col]
       return df_new

    def entropy_weights(df_feature, indicators):
       scaler = MinMaxScaler()
       score_matrix = scaler.fit_transform(df_feature[indicators])
       P = score_matrix / score_matrix.sum(axis=0)
       E = -np.nansum(P * np.log(P + 1e-12), axis=0) / np.log(P.shape[0])
       d = 1 - E
       w = d / d.sum()
       return pd.Series(w, index=indicators)

 # === Part 5: TOPSIS + 可视化权重
    score_cols = ["日均消费金额", "总消费金额", "单次消费均值", "最小单次消费", "最大单次消费"]
    score_directions = ["负向"] * len(score_cols)
    features_transformed = transform_indicators(features[["校园卡号"] + score_cols], score_cols, score_directions)
    weights = entropy_weights(features_transformed, score_cols)
     print(weights.sort_values(ascending=False).round(4))

    weights.sort_values().plot(kind="barh", color="teal")
    plt.title("各指标的熵权法权重分布")
```

```
    plt.xlabel("权重值")
     plt.tight_layout()
     plt.savefig("图 3_熵权法权重分布图.png", dpi=600)
    plt.show()"

全体学生综合贫困得分直方图
 # === Part 6: TOPSIS 评分函数
    def entropy_topsis(df_feature, id_col, indicators):
        scaler = MinMaxScaler()
        score_matrix = scaler.fit_transform(df_feature[indicators])
        P = score_matrix / score_matrix.sum(axis=0)
        E = -np.nansum(P * np.log(P + 1e-12), axis=0) / np.log(P.shape[0])
        d = 1 - E
        w = d / d.sum()
        ideal_best = score_matrix.max(axis=0)
        ideal_worst = score_matrix.min(axis=0)
        D_plus = np.sqrt(((score_matrix - ideal_best) ** 2 * w).sum(axis=1))
        D_minus = np.sqrt(((score_matrix - ideal_worst) ** 2 * w).sum(axis=1))
        C = D_minus / (D_plus + D_minus)
        result = df_feature[[id_col]].copy()
        result["综合得分"] = C
        result["困难排序"] = result["综合得分"].rank(ascending=False, method="min")
       return result

    # TOP50 筛选
    topsis_result = entropy_topsis(features_transformed, "校园卡号", score_cols)
    top50 = topsis_result.sort_values(by="综合得分").head(50).merge(features, on="校园卡号")
# === 全体学生的 TOPSIS 综合得分直方图 ===
   plt.figure(figsize=(10, 4))
   histplot = sns.histplot(topsis_result["综合得分"], bins=30, kde=True, color="steelblue")
   for patch in histplot.patches:
        height = patch.get_height()
        if height > 0:
            plt.text(patch.get_x() + patch.get_width()/2, height + 1, f"{int(height)}",
                     ha='center', va='bottom', fontsize=10)
    plt.xlabel("综合贫困得分 $C_i$")
    plt.ylabel("学生人数")
    plt.title("全体学生综合贫困得分直方图（熵权-TOPSIS）", fontsize=12)
    plt.grid(axis="y", linestyle="--", alpha=0.5)
    plt.tight_layout()
    plt.savefig("图 4_全体学生综合贫困得分直方图.png", dpi=600)
     plt.show()
 # === # === 5.2 问题二：困难识别与 TOPSIS 评分：基于整理后的“指标数据”使用熵权法评分
```

```
# 说明：以下指标为同伴从原始数据中提炼，已包含合理建模特征
# 读取第 5.1 节提取并保存的关键画像变量（由原始数据生成）
indicator_df = pd.read_excel("指标数据.xlsx")  # 由原始数据生成，仅作中间数据载入
# 指标与方向列表
indicators = [
"单次平均消费金额(反向指标）", "消费频次(正向指标)", "异常低消比例(正向)",
"场所多样性(反向)", "核心场所消费比(正向)", "日均消费(反向)", "消费金额波动(反向)"
]
directions = ["负向", "正向", "正向", "负向", "正向", "负向", "负向"]
# 正向化函数
def transform_indicators(df, indicators, directions):
    df_new = df.copy()
    for col, direction in zip(indicators, directions):
        if direction == '负向':
            max_val, min_val = df[col].max(), df[col].min()
            df_new[col] = max_val + min_val - df[col]
    return df_new

# 熵权法函数
def entropy_weights(df_feature, indicators):
    scaler = MinMaxScaler()
    score_matrix = scaler.fit_transform(df_feature[indicators])
    P = score_matrix / score_matrix.sum(axis=0)
    E = -np.nansum(P * np.log(P + 1e-12), axis=0) / np.log(P.shape[0])
    d = 1 - E
    w = d / d.sum()
    return pd.Series(w, index=indicators)

# 得分计算函数
def entropy_topsis(df_feature, id_col, indicators):
    scaler = MinMaxScaler()
    score_matrix = scaler.fit_transform(df_feature[indicators])
    P = score_matrix / score_matrix.sum(axis=0)
    E = -np.nansum(P * np.log(P + 1e-12), axis=0) / np.log(P.shape[0])
    d = 1 - E
    w = d / d.sum()
    ideal_best = score_matrix.max(axis=0)
    ideal_worst = score_matrix.min(axis=0)
    D_plus = np.sqrt(((score_matrix - ideal_best) ** 2 * w).sum(axis=1))
    D_minus = np.sqrt(((score_matrix - ideal_worst) ** 2 * w).sum(axis=1))
    C = D_minus / (D_plus + D_minus)
    result = df_feature[[id_col]].copy()
    result["综合得分"] = C
    result["困难排序"] = result["综合得分"].rank(ascending=False, method="min")
```

```
        return result
    # 正向化 + 权重输出 + TOPSIS 评分
    transformed = transform_indicators(indicator_df, indicators, directions)
    权重结果 = entropy_weights(transformed, indicators).sort_values(ascending=False).round(4)
    评分结果 = entropy_topsis(transformed, "校园卡号", indicators)
    权重结果, 评分结果.head()
```

附录 2

问题三：Python 代码

```
# -*- coding: utf-8 -*-
"""
校园精准资助：从消费数据分析、异常识别、评分建模到资助等级划分
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import rcParams
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, calinski_harabasz_score
import math

# 设置全局字体为宋体
rcParams['font.family'] = 'Microsoft YaHei'
plt.rcParams['axes.unicode_minus'] = False

# 一、加载清洗后数据
cleaned_data = pd.read_excel("1.xlsx")

# 二、数据预处理
cleaned_data = cleaned_data[cleaned_data['消费类型'] == '消费']
cleaned_data = cleaned_data.dropna(subset=['消费金额（元）'])
cleaned_data['消费时间'] = pd.to_datetime(cleaned_data['消费时间'])
cleaned_data.sort_values(by=['校园卡号', '消费时间'], inplace=True)

# 三、标记核心场所
core_places = ['第一食堂', '第二食堂', '第三食堂', '第四食堂', '第五食堂', '好利来食品店']
cleaned_data['场所类型'] = cleaned_data['消费地点'].apply(lambda x: '核心场所' if x in
core_places else '辅助场所')

# 四、构建学生画像
def get_student_features(df):
    results = []
    for sid, group in df.groupby('校园卡号'):
        gender = group['性别'].iloc[0]
        total = group['消费金额（元）'].sum()
        count = group.shape[0]
        avg_once = total / count if count > 0 else 0
```

```
            daily = total / 30
            core_group = group[group['场所类型'] == '核心场所']
            core_ratio = len(core_group) / count if count > 0 else 0
            low_threshold = np.percentile(core_group['消费金额（元）'], 5) if len(core_group) > 0
else 0
            low_consume = (core_group['消费金额（元）'] < low_threshold).sum()
            low_ratio = low_consume / len(core_group) if len(core_group) > 0 else 0
            diversity = group['消费地点'].nunique()
            std = group['消费金额（元）'].std()
            cv = std / avg_once if avg_once > 0 else 0
            results.append([sid, gender, avg_once, count, low_ratio, diversity, core_ratio, daily, cv])
    return pd.DataFrame(results, columns=['校园卡号', '性别', '单次平均消费金额', '消费频次',
'异常低消比例', '场所多样性', '核心场所消费比', '日均消费', '消费金额波动'])

student_features = get_student_features(cleaned_data)

# 五、可视化：单次平均消费金额直方图
plt.figure(figsize=(8, 5))
counts, bins, patches = plt.hist(student_features['单次平均消费金额'], bins=20, color='skyblue',
edgecolor='black')
plt.title('单次平均消费金额分布')
plt.xlabel('元')
plt.ylabel('学生数量')
for count, patch in zip(counts, patches):
    plt.text(patch.get_x() + patch.get_width()/2, count, int(count), ha='center', va='bottom',
fontsize=8)
plt.tight_layout()
plt.savefig("图 1_单次平均消费金额分布.png", dpi=600)

# 六、构建评价指标体系并归一化
raw_scores = student_features.set_index('校园卡号')[['单次平均消费金额', '消费频次', '异常低
消比例', '场所多样性', '核心场所消费比', '日均消费', '消费金额波动']].copy()
reverse_cols = ['单次平均消费金额', '场所多样性', '日均消费', '消费金额波动']
for col in reverse_cols:
    raw_scores[col] = raw_scores[col].max() - raw_scores[col]

scaler = MinMaxScaler()
normalized = pd.DataFrame(scaler.fit_transform(raw_scores), index=raw_scores.index,
columns=raw_scores.columns)

# 七、计算指标相关性热力图
plt.figure(figsize=(8, 6))
corr = raw_scores.corr()
sns.heatmap(corr, annot=True, cmap='RdYlBu_r', fmt=".2f")
```

```
plt.title("评分指标间的相关性热力图")
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.savefig("图 2_评分指标相关性热力图.png", dpi=600)

# 八、熵权法求权重
k = 1.0 / math.log(len(normalized))
P = normalized / normalized.sum()
E = -k * (P * np.log(P + 1e-8)).sum()
d = 1 - E
w = d / d.sum()

# 可视化权重柱状图
plt.figure(figsize=(8, 5))
bars = plt.bar(w.index, w.values)
plt.title("熵权法下各指标权重")
plt.ylabel("权重值")
plt.xticks(rotation=30)
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval, f"{yval:.2f}", ha='center', va='bottom',
fontsize=8)
plt.tight_layout()
plt.savefig("图 3_熵权指标权重柱状图.png", dpi=600)

# 九、TOPSIS 综合得分
Z = normalized * w
A_plus = Z.max()
A_minus = Z.min()
D_plus = np.linalg.norm(Z - A_plus, axis=1)
D_minus = np.linalg.norm(Z - A_minus, axis=1)
C = D_minus / (D_plus + D_minus)
raw_scores['贫困得分'] = C

# 可视化得分分布
plt.figure(figsize=(8, 5))
counts, bins, patches = plt.hist(C, bins=20, color='lightgreen', edgecolor='black')
plt.title('TOPSIS 综合贫困得分分布')
plt.xlabel('贫困得分')
plt.ylabel('学生数量')
for count, patch in zip(counts, patches):
    plt.text(patch.get_x() + patch.get_width()/2, count, int(count), ha='center', va='bottom',
fontsize=8)
```

```
plt.tight_layout()
plt.savefig("图 4_TOPSIS 贫困得分分布图.png", dpi=600)

# 十、K-means 聚类评估
selected = raw_scores.sort_values(by='贫困得分').head(50)
X = selected.drop(columns='贫困得分')
k_range = range(2, 8)
sse, sils, chs = [], [], []
for k in k_range:
    model = KMeans(n_clusters=k, random_state=0).fit(X)
    sse.append(model.inertia_)
    sils.append(silhouette_score(X, model.labels_))
    chs.append(calinski_harabasz_score(X, model.labels_))

# 肘部法则
plt.figure(figsize=(8, 5))
plt.plot(k_range, sse, marker='o')
for i, val in enumerate(sse):
    plt.text(k_range[i], val, f"{val:.0f}", ha='center', va='bottom', fontsize=8)
plt.title("肘部法则：聚类数与 SSE")
plt.xlabel("K 值")
plt.ylabel("SSE")
plt.tight_layout()
plt.savefig("图 5_肘部法则_SSE 曲线.png", dpi=600)

# 轮廓系数法
plt.figure(figsize=(8, 5))
plt.plot(k_range, sils, marker='^')
for i, val in enumerate(sils):
    plt.text(k_range[i], val, f"{val:.2f}", ha='center', va='bottom', fontsize=8)
plt.title("轮廓系数法：聚类数与轮廓系数")
plt.xlabel("K 值")
plt.ylabel("轮廓系数")
plt.tight_layout()
plt.savefig("图 6_轮廓系数法评估曲线.png", dpi=600)

# CH 法
plt.figure(figsize=(8, 5))
plt.plot(k_range, chs, marker='s')
for i, val in enumerate(chs):
    plt.text(k_range[i], val, f"{val:.0f}", ha='center', va='bottom', fontsize=8)
plt.title("CH 指数法：聚类数与 CH 值")
plt.xlabel("K 值")
plt.ylabel("CH 值")
```

```
plt.tight_layout()
plt.savefig("图 7_CH 值评估曲线.png", dpi=600)

# 十一、最终聚类 + 可视化
kmeans = KMeans(n_clusters=4, random_state=0).fit(X)
selected['资助等级'] = kmeans.labels_ + 1

plt.figure(figsize=(8, 6))
sns.scatterplot(x=selected['消费频次'], y=selected['单次平均消费金额'], hue=selected['资助等级
'], palette='Set2')
plt.title('K-means 聚类下的资助等级分布')
plt.xlabel('消费频次')
plt.ylabel('单次平均消费金额')
plt.tight_layout()
plt.savefig("图 8_Kmeans 聚类等级散点图.png", dpi=600)

plt.figure(figsize=(8, 5))
sns.boxplot(data=selected, x='资助等级', y='日均消费')
plt.title("不同资助等级学生的日均消费对比")
plt.xlabel("资助等级")
plt.ylabel("日均消费金额")
plt.tight_layout()
plt.savefig("图 9_各等级日均消费箱型图.png", dpi=600)

# 输出 Excel
to_save = selected.sort_values(by='资助等级')
to_save.to_excel("资助学生名单与等级.xlsx")

# 十二、金额区间测算
def compute_grant_ranges_refined(df, value_col='日均消费', level_col='资助等级'):
    base_days = 150
    grant_list = []

    # 找出 I 类特困组的中位数，作为基础线
    base_group = df[df[level_col] == 1][value_col]
    base_median = base_group.median()
    base_need = base_median * base_days
    base_grant = base_need * 1.10   # 10%补贴

    # 补贴比例映射
    subsidy_ratios = {1: 1.10, 2: 1.00, 3: 0.90, 4: 0.80, 5: 0.70, 6: 0.60}

    for level in sorted(df[level_col].unique()):
        group = df[df[level_col] == level]
```

```
            median_daily = group[value_col].median()
            est_need = median_daily * base_days
            subsidy = subsidy_ratios.get(level, 0.60)
            suggested = base_need * subsidy
            low, high = round(suggested - 50, 2), round(suggested + 50, 2)

            grant_list.append({
                '资助等级': level,
                '人数': len(group),
                '日均消费中位数': round(median_daily, 2),
                '估算学期消费': round(est_need, 2),
                '建议资助金额': round(suggested, 2),
                '建议资助区间': f'{low} ~ {high} 元'
            })

    return pd.DataFrame(grant_list)

# 调用区间测算
grant_ranges_df = compute_grant_ranges_refined(selected)
grant_ranges_df.to_excel("不同等级资助金额区间测算表.xlsx", index=False)

# 可视化一：各等级建议金额对比
plt.figure(figsize=(8, 5))
bars = plt.bar(grant_ranges_df['资助等级'], grant_ranges_df['建议资助金额'], color='skyblue',
edgecolor='black')
plt.title('各资助等级建议资助金额')
plt.xlabel('资助等级')
plt.ylabel('金额（元）')
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval + 20, f"{int(yval)}", ha='center', fontsize=9)
plt.tight_layout()
plt.savefig("图 10_各等级建议资助金额柱状图.png", dpi=600)

# 可视化二：预算总金额计算
grant_ranges_df['预算总额'] = grant_ranges_df['建议资助金额'] * grant_ranges_df['人数']
plt.figure(figsize=(8, 5))
bars2 = plt.bar(grant_ranges_df['资助等级'], grant_ranges_df['预算总额'], color='lightgreen',
edgecolor='black')
plt.title('各等级资助预算总额')
plt.xlabel('资助等级')
plt.ylabel('总金额（元）')
for bar in bars2:
    yval = bar.get_height()
```

```
        plt.text(bar.get_x() + bar.get_width() / 2, yval + 500, f"{int(yval)}", ha='center', fontsize=9)
plt.tight_layout()
plt.savefig("图 11_各等级预算总额柱状图.png", dpi=600)

# 汇总总预算输出
total_budget = grant_ranges_df['预算总额'].sum()
print(f"预计总资助预算：{round(total_budget, 2)} 元")



# 十三、针对第一问的消费趋势分析
# 图 12：每日消费总额趋势图
cleaned_data['日期'] = cleaned_data['消费时间'].dt.date
by_day = cleaned_data.groupby('日期')['消费金额（元）'].sum()

plt.figure(figsize=(10, 5))
by_day.plot(marker='o', linestyle='-')
plt.title("每日消费总额趋势图")
plt.xlabel("日期")
plt.ylabel("消费总金额（元）")
plt.grid(True, linestyle='--', alpha=0.5)
plt.tight_layout()
plt.savefig("图 12_每日消费趋势图.png", dpi=600)

# 图 13：按小时消费频率直方图
cleaned_data['小时'] = cleaned_data['消费时间'].dt.hour
hour_counts = cleaned_data['小时'].value_counts().sort_index()

plt.figure(figsize=(10, 5))
bars = plt.bar(hour_counts.index, hour_counts.values, color='orange', edgecolor='black')
plt.title("各时段消费频率分布图")
plt.xlabel("小时（0-23）")
plt.ylabel("消费记录数")
plt.xticks(range(0, 24))
for i, bar in enumerate(bars):
        yval = bar.get_height()
        plt.text(bar.get_x() + bar.get_width()/2, yval + 5, int(yval), ha='center', fontsize=9)
plt.tight_layout()
plt.savefig("图 13_小时消费频率分布.png", dpi=600)

# 图 14：一周内消费频率分布
cleaned_data['星期'] = cleaned_data['消费时间'].dt.dayofweek
weekday_map = ['周一', '周二', '周三', '周四', '周五', '周六', '周日']
cleaned_data['星期中文'] = cleaned_data['星期'].map(dict(enumerate(weekday_map)))
weekday_counts = cleaned_data['星期中文'].value_counts().reindex(weekday_map)
```

```
plt.figure(figsize=(8, 5))
bars2 = plt.bar(weekday_counts.index, weekday_counts.values, color='steelblue', edgecolor='black')
plt.title("周内消费频率分布图")
plt.xlabel("星期")
plt.ylabel("消费次数")
for bar in bars2:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval + 5, int(yval), ha='center', fontsize=9)
plt.tight_layout()
plt.savefig("图 14_周内消费频率分布.png", dpi=600)
```