|  |  |  |
| --- | --- | --- |
| 选题 | **2024年第十四届APMCM**  **亚太地区大学生数学建模竞赛（中文赛项）** | 参赛编号 |
| B | apmcm\*\*\*\*\*\* |

标题（此处换成论文的标题）

摘要

1. 问题重述

1.1问题背景

洪水作为一种严重的自然灾害，对人类社会和生态环境造成了巨大的影响。历史上，洪水灾害的发生往往与自然因素如暴雨、融冰化雪、风暴潮等有关。然而，随着人口的增长和人类活动的增加，如乱砍滥伐、围湖造田等，这些活动加剧了地表状态的改变，进而影响了汇流条件和洪水灾害的程度。近年来，全球洪水灾害频发，造成了巨大的经济损失和人员伤亡。因此，对洪水灾害进行数据分析与预测，对于提前预警和减少灾害损失具有重要意义。

1.2问题要求

附件数据给出了各地季风强度、地形排水、河流管理等基本信息，以及该地发生洪水的概率。为了能更好地根据已知信息预测洪水的发生概率，最小化灾害损失，现需结合实际情况与所给信息建立数学模型，分析以下问题：

问题1：依据附件1中给出的基本信息，可视化处理相关数据，并单独分析洪水发生概率与每个基本信息的相关性，判断其关联程度大小，给出相关结论和建议。

问题2：将train.csv中洪水发生的概率聚类成不同风险类别，分析高、中、低风险洪水事件的指标特征。选取合适的指标，计算不同指标的权重，建立预警评价模型，并进行模型的灵敏度分析。

问题3：基于问题1中指标分析的结果，建立洪水发生概率的预测模型。从20个指标中选取合适指标，预测洪水发生的概率，并验证模型的准确性。探讨如果仅用5个关键指标，如何调整改进预测模型。

问题4：利用问题2中建立的洪水发生概率预测模型，预测附件test.csv中所有事件的洪水发生概率，并将结果填入submit.csv。绘制预测概率的直方图和折线图，分析其分布特征，判断是否服从正态分布。

附件：

1. train.csv - 包含超过100万条洪水数据，涵盖洪水事件id和20个指标得分，以及发生洪水的概率。
2. test.csv - 包含超过70万条洪水数据，涵盖洪水事件id和20个指标得分，缺少发生洪水的概率。
3. submit.csv - 包含test.csv中的洪水事件id，缺少发生洪水的概率。
4. 问题分析

**2.1问题1的分析**

针对问题1，首先对train.csv中所给数据进行最大值、最小值、平均值、中位数等基本统计量计算、数据预处理等基础性工作,并记录洪水概率的最大、最小值，基于这些数据，采用概率密度直方图、核密度估计（Kernel Density Estimation, KDE）图可视化洪水概率的分布情况，同时用频数统计直方图可视化train.csv与test.csv中季风强度、地形排水等各个指标特征的分布情况，并计算上述所有指标分布的偏度与峰度系数用以量化各项评估指标与洪水概率的分布情况。随后，引入Spearman相关系数进相关性分析，并加以假设检验，用以量化各项指标与洪水概率之间的关系。最后，综合以上获得的相关性、数据分布情况等分析结果进行统计分析，指出对洪水概率影响较大的指标，并据此提出相应的预防措施。

**2.2问题2的分析**

针对问题2，首先标准化train.csv中洪水概率数据，并利用K-Means++聚类算法将其聚类为三簇，随后对聚类结果进行箱型图、频数分布直方图可视化，基于簇进行数据表分组，并计算每个分组的最大、最小值，用以评估聚类结果是是否存在交错，最后划分风险为高、中、低三类，将其作为分类标签，并对其分别进行统计分析可视化，用以分析不同风险对应指标特征。风险预警评价问题实质是基于连续数值到逻辑类别区间映射的回归+分类的合并问题，首先基于先前得出的三类分组数据中最大、最小值构造连续回归数值转离散类别数值的映射，并训练逻辑回归模型建立符合分布的洪水概率到风险等级的连续映射，实现有空缺映射区间的填充工作，以保证最终回归+分类模型评价能力的鲁棒性，随后，基于所有指标构建线性回归模型作为基准模型，实现较优结果，探索指标特征与洪水概率间潜在的线性相关性，并对所有指标之和与洪水概率进行可视化以确认其线性相关性，因此引入指标和作为关键指标，然后将季风强度、地形排水等指标在样本维度上从小到大排序，构造排序指标，从而消除指标维度上的数据差异性，最后，将排序指标与指标和作为最终指标训练CatBoost模型实现了优于基准模型的效果，并且其在三分类任务上同样表现优秀，同时，利用K折交叉验证法验证了模型的低灵敏度与高泛化性。

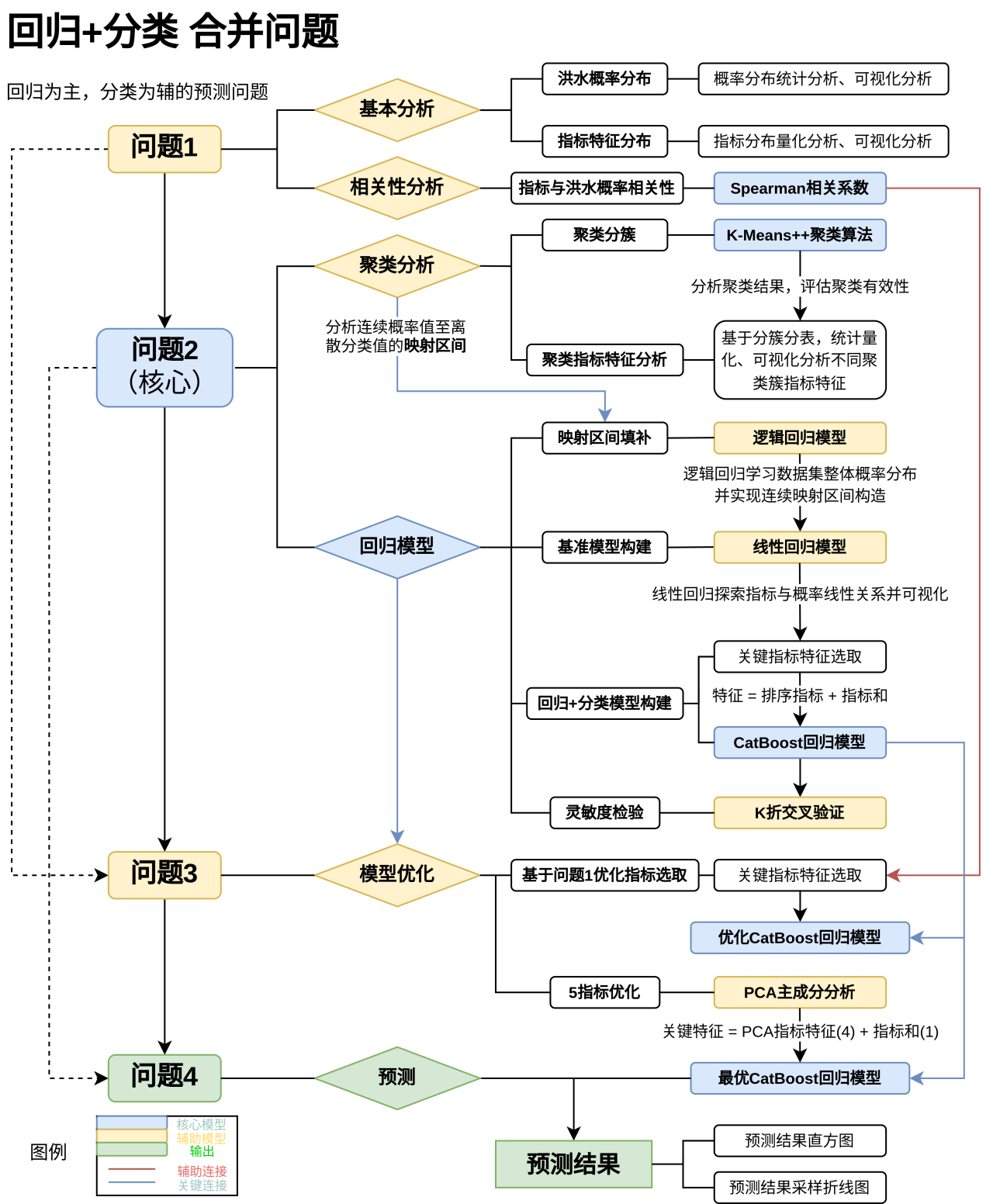
**2.3问题3的分析**

针对问题3，首先基于问题1分析出的各指标与洪水概率的Spearman相关系数选出相关性较高的指标特征作为基本指标，并加入指标和作为关键指标，训练CatBoost回归模型，得出较优回归预测模型。随后，使用主成分分析（Principle Component Analysis，PCA）进行数据降维，将20维指标特征降至4维作为关键指标，同时依旧加入指标和组成5个指标，用以训练CatBoost回归模型，得出最优回归预测模型。

**2.4问题4的分析**

针对问题4，直接选用问题3中得出的最优CatBoost回归模型进行test.csv的特征工程，才用最优5个关键指标进行结果预测，于submit.csv中填写预测结果，并对预测结果使用直方图、折线图（随机采样1000样本）进行可视化分析，计算其偏度峰度系数，分析其分布正态性，并与其他两个性能较低的CatBoost回归模型进行分布对比，验证其预测结果正态性与合理性。

**文章总体思路如图1所示：**

****

**图1总体思路图**

1. 模型假设

四、符号说明

五、数据侧写

1. 问题1模型的建立与求解

**6.1基于数据的数字特征进行描述**

题目要求对数据进行可视化处理，为了更好地描述洪水发生概率分布、基本信息特征的分布规律，本文引入峰度系数、偏度系数来描述统计数据。

1. **峰度系数**

峰度系数（Kurtosis）用于衡量数据分布的峰度程度。例如时，数据的分布为正态分布；时，数据分布的峰度较小，数据更分散；时，数据分布的峰度较大，数据更集中。并且越大，数据分布的图像更加“陡峭”。峰度系数的计算公式如下：



1. **偏度系数**

偏度系数（Skewness）用于衡量数据分布的偏斜程度。例如时，数据的分布为正态分布；时，数据呈左偏分布；时，数据呈右偏分布。并且越大，数据分布的偏斜越明显。峰度系数的计算公式如下：



**6.2数据可视化**

利用python作出洪水发生概率密度图像、基本信息的分布柱状图如下：

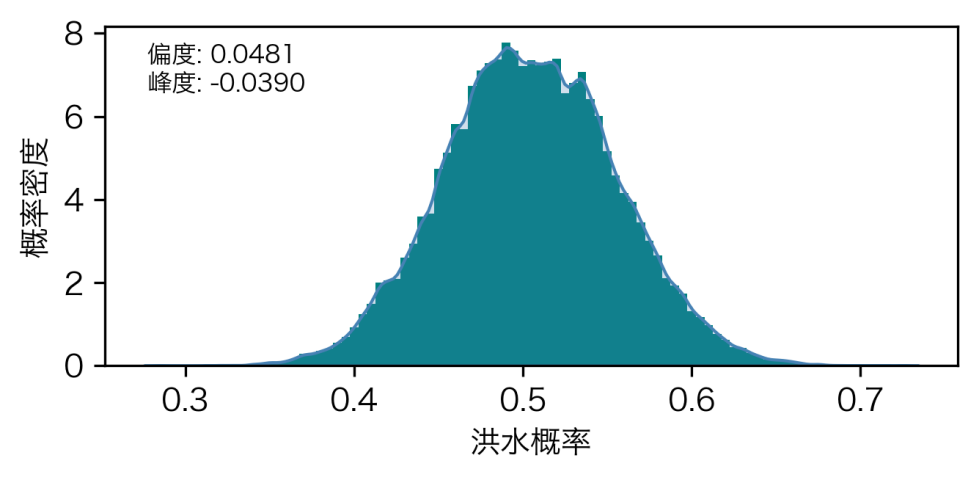
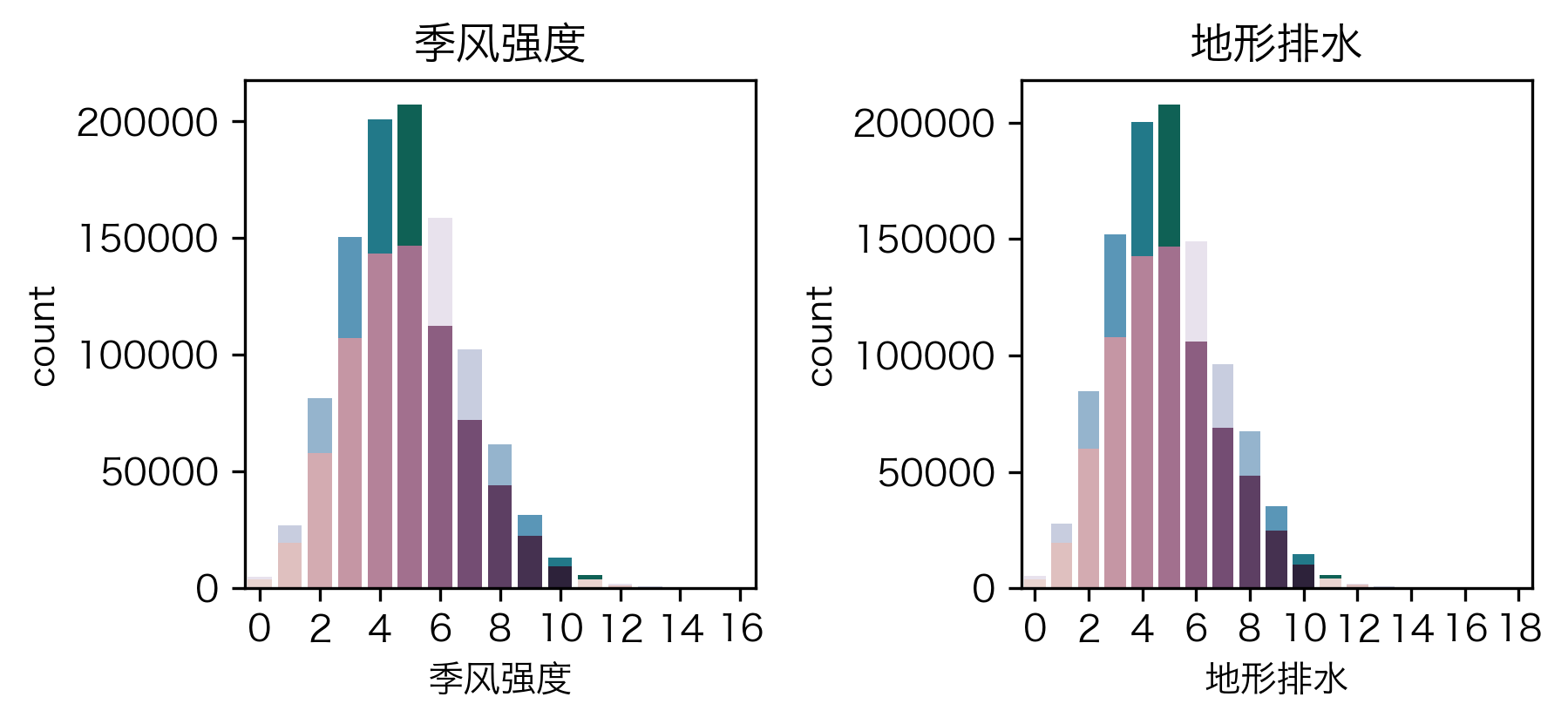


图2洪水概率密度的图像



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 训练集偏度系数 | 训练集峰度系数 | 测试集偏度系数 | 测试集峰度系数 |
| 平均值 | 0.441813837 | 0.247223868 | 0.44093833 | 0.247306689 |
| 标准差 | 0.011392705 | 0.035052238 | 0.012480602 | 0.034875173 |
| 最小值 | 0.419867945 | 0.187350475 | 0.413857617 | 0.182297125 |
| 25% | 0.436629875 | 0.232680673 | 0.435625489 | 0.228729943 |
| 50% | 0.441398792 | 0.243670623 | 0.442318853 | 0.246637074 |
| 75% | 0.449001103 | 0.262154138 | 0.449952157 | 0.263493856 |
| 最大值 | 0.464098419 | 0.339472861 | 0.460378327 | 0.315153435 |

**6.1利用Spearman相关系数判断指标与洪水概率之间的相关性**

Spearman相关系数是一种非参数的相关性度量，可以等级化变量之间相关性，并且不局限于线性关系，适用于非线性关系的评估，其计算公式如下：

**[]**

其中，[]与[]均为对应观测值第i个取值的等级，[]与[]均为对应观测值取值的平均等级，N为观测值的总数量，[]

Spearman相关系数评估了两变量之间的单调关系，即两个变量同时增加（或减少）时相关系数趋近于1（或-1）；两个变量的变化之间未出现明显关系时趋近于0。

利用Python求解Spearman相关系数，结果如表1所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Spearman相关系数** | | | | |
| 基础设  施恶化 | 地形  排水 | 季风  强度 | 大坝  质量 | 河流  管理 |
| 0.18139 | 0.18048 | 0.18028 | 0.17945 | 0.17893 |
| 淤积 | 人口  得分 | 气候  变化 | 森林  砍伐 | 滑坡 |
| 0.17873 | 0.17799 | 0.17711 | 0.17701 | 0.17663 |
| 海岸脆  弱性 | 农业  实践 | 湿地  损失 | 流域 | 政策  原因 |
| 0.17029 | 0.17514 | 0.17513 | 0.17464 | 0.17387 |
| 规划  不足 | 城市化 | 侵蚀 | 排水  系统 | 无效  防灾 |
| 0.17325 | 0.17265 | 0.17143 | 0.17044 | 0.17597 |
| 规划  不足 | 城市化 | 侵蚀 | 排水  系统 | 无效  防灾 |
| 0.18139 | 0.18048 | 0.18028 | 0.17945 | 0.17893 |

**表 1 指标与洪水概率的Spearman相关系数表**

由表 1 可知：

1. 基础设施恶化对洪水的影响最大，海岸脆弱性对洪水的影响最小。
2. 20个指标与洪水概率的Spearman相关系数都在0.175左右，说明20个指标均与洪水发生概率呈正相关，但是相关性均较弱。

**6.2 Spearman相关系数的检验**

本文对20个指标采取 Spearman 相关系数检验，检验步骤如下：

**(1)提出假设**

原假设𝐻0：Spearman 系数 ≠ 0

备择假设𝐻1：Spearman 系数 = 0

设定置信水平为 99.5%

1. **计算P值**

本文采用Python进行spearman相关系数检验，结果如表2所示

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Spearman相关系数检验的P值 | | | | | |
|  | 季风强度 | 地形排水 | 河流管理 | ... | 政策因素 |
| 季风强度 | 1.0000 | 0.0000 | 0.0000 | ... | 0.0000 |
| 地形排水 | 0.0000 | 1.0000 | 0.0000 | ... | 0.0000 |
| 河流管理 | 0.0000 | 0.0000 | 1.0000 | ... | 0.0000 |
| ... | ... | ... | ... | ... | ... |
| 政策因素 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 |

**表 2 Spearman相关系数表的P值**

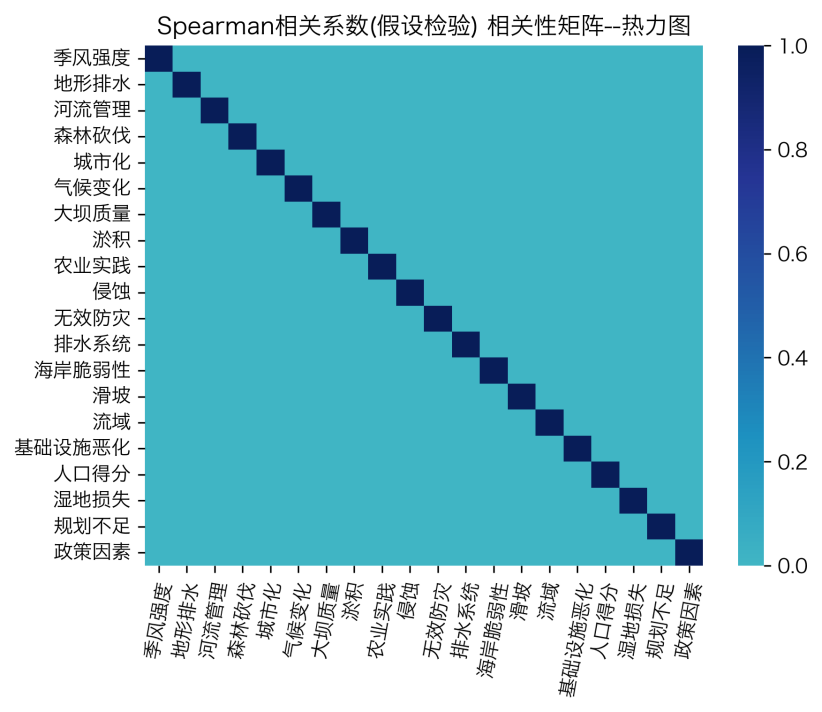


表2是Spearman相关系数检验的P值构成的矩阵，其对角线元素为1，其余元素均为0，即其为单位矩阵，可知任意两指标之间的P值均为0，故接受原假设，认为Spearman系数≠0。

七、问题2模型的建立与求解

7.1低，中，高风险聚类

**7.1.1.K-Means++聚类过程**

K-Means++是K-Means聚类算法的一种改进版本，它通过智能地选择初始聚类中心来提高K-Means的性能。K-Means++的目标是更好地选择初始聚类中心，以减少算法的迭代次数，提高聚类的质量。具体步骤如下：

1. **选择第一个聚类中心**：随机选择一个数据点作为第一个聚类中心。
2. **选择后续聚类中心：**对于每个数据点，计算它与已选择的聚类中心之间的最短距离D(x)（即与最近的聚类中心之间的距离）。每个数据点作为下一个聚类中心的概率为:

其中x代表任意一个数据点，X表示数据集。这样保证了距离更远的数据点更有可能被选为下一个聚类中心，以确保新的聚类中心较好地覆盖数据分布。

1. **完成聚类中心的选择：**重复步骤2直到选择足够数量的聚类中心（一般为k个，我们称将数据分成了k个簇）。
2. **计算每个数据点到聚类中心的距离**：对于每个数据点，计算它与每个聚类中心的距离，通常使用欧氏距离或其他距离度量方式。对于数据点与聚类中心的距离由以下公式给出：

]

其中为第i个样本点，为第j个聚类中心，是第i个样本点的第k个分量，是第j个聚类中心的第k个分量。

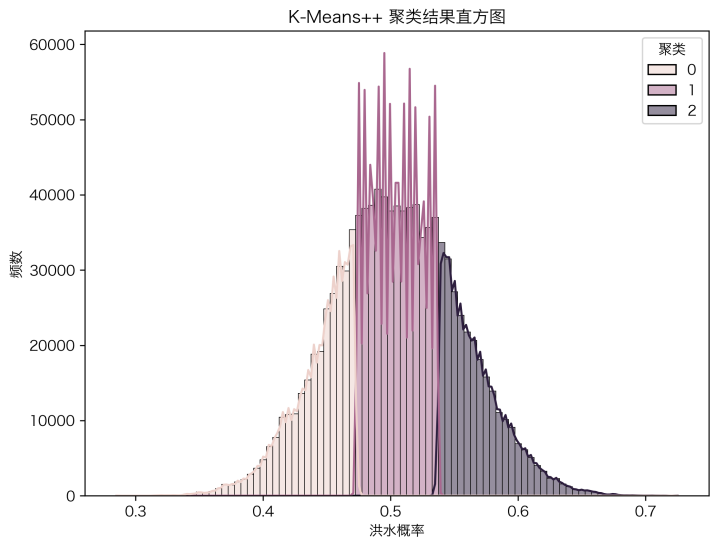
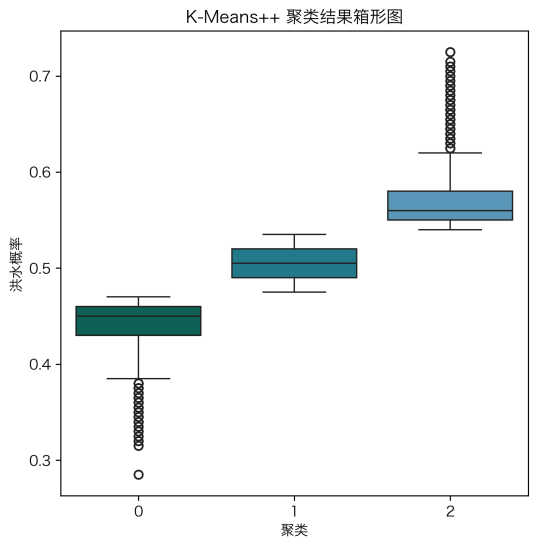
1. **分配数据点到最近的聚类中心：**将每个数据点分配到距离它最近的聚类中心所属的簇。
2. **更新聚类中心：**重新计算每个簇的中心，即该簇所有数据点的平均值。将该平均值作为新的聚类中心。即新的中心为：

其中为新的聚类中心，其中为以为聚类中心的聚类，为中数据点的个数。

1. **完成聚类**：重复步骤4，5，6，直到聚类中心不再改变，或达到预定义的停止条件（比如达到最大迭代次数）。

最终得到每个数据点所属的簇，即完成了K-Means++聚类。

使用Python进行求解，再用箱型图和直方图进行可视化，得到以下结果：



从图中不难看出，低，中，高风险的洪水概率是无交错的且大致服从正态分布，但是对于风险评级的划分，我们希望得到[0,1]区间上的三个子区间，而事实上我们K-Means++得到的三个区间分别为[0.285,0.47]，[0.475,0.535]，[0.54,0.725]，显然这些区间是存在空缺的，为了填补这个空缺实现一个完整的回归区间划分使其较好地满足回归+三分类模型，基于对train.csv中洪水概率的统计分析，我们使用逻辑回归对其进行区间填补。

**7.1.2 逻辑回归填补区间**

逻辑回归,即对数概率回归，是一种线性模型，是一种用于分类问题的算法。通过学习特征与类别之间的关系，逻辑回归可以预测新数据点属于哪个类别，输出类别的概率，其过程如下：

1. **收集数据**：收集具有标签的训练数据集，每个数据点包括特征值和类别标签。
2. **特征工程**：对数据进行特征提取和选择，包括数据清洗、特征缩放、特征选择等操作。
3. **定义预测函数**：假设我们的模型可以用如下形式表达：

其中[]是Sigmoid函数，定义为：

[]

的导数满足]。

则最终预测函数表达式为：

其中，为对应权重,需要后续过程计算得出。

最后样本为正样本的概率为，样本为负样本的概率为[。

1. **定义损失函数**：我们使用交叉熵损失函数来度量模型预测值与真实标签之间的误差（由于篇幅原因，我们不给出推导过程）：
2. **最小化损失函数：**通过梯度下降等优化算法，最小化损失函数，得到最佳的参数。
3. **预测：**当模型训练完成后，我们可以用训练好的参数对新数据进行分类预测。分类规则为：如果正样本的概率大于负样本的概率，则样本被判定为正样本，否则被判定为负样本，这等价于

首先将训练集给出的洪水概率作为特征，将聚类出的类别作为分类标签，训练逻辑回归分类器，原始的区间[0.2825, 0.7275]微分切片为1048575份，为了提升精确度，将其提升100倍，生成一个相对连续的概率区间划分序列，随后将该序列用训练出的逻辑回归分类模型进行分类，成功找到三个基本连续的区间[0.282500, 0.472481], [0.472481, 0.537549], [0.537549, 0.727500],事实上，这三个区间并不是严格相邻，但是，此时区间之间的距离很小，基本视为连续。因此，得出区间划分的结论为：

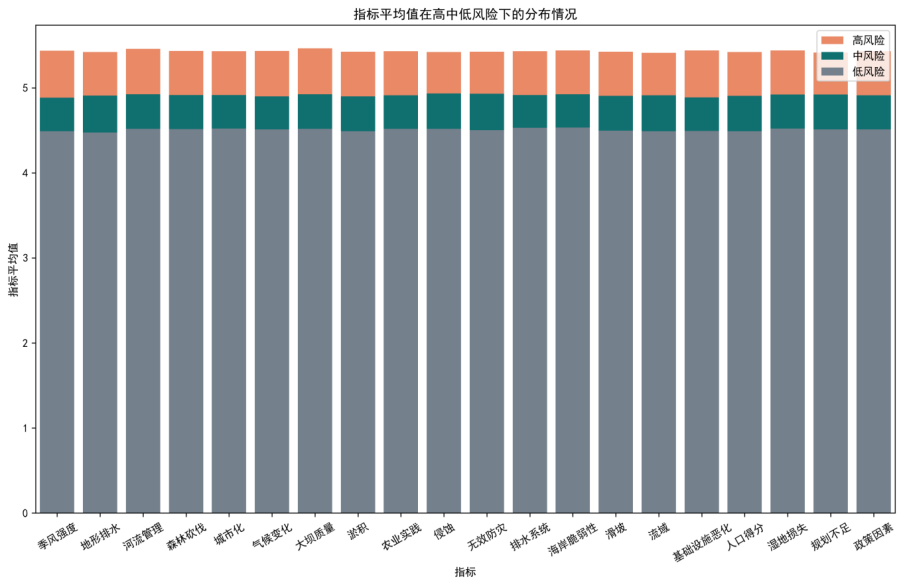
至此，我们将所有train.csv中所有数据根据洪水发生的概率划分成了低，中，高风险三个类别，接下来分析三个不同风险类别的指标特征。

**7.2分析具有高、中、低风险的洪水事件的指标特征**

对高、中、低风险的洪水事件进行统计分析以及可视化，得如下结果（以低风险为例，其余见附表）：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 低风险区的各指标分布情况 | | | | | |
|  | **季风强度** | **地形排水** | **河流管理** | **森林砍伐** | **城市化** |
| 平均值 | 4.489 | 4.474 | 4.518 | 4.514 | 4.520 |
| 最小值 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 18 | 16 | 16 | 16 |
|  | **气候变化** | **大坝质量** | **淤积** | **农业实践** | **侵蚀** |
| 平均值 | 4.511 | 4.517 | 4.491 | 4.518 | 4.517 |
| 最小值 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 17 | 16 | 16 | 16 | 17 |
|  | **无效防灾** | **排水系统** | **海岸**  **脆弱性** | **滑坡** | **流域** |
| 平均值 | 4.503 | 4.532 | 4.535 | 4.496 | 4.491 |
| 最小值 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 17 | 17 | 16 | 16 |
|  | **基础**  **设施恶化** | **人口得分** | **湿地损失** | **规划不足** | **政策因素** |
| 平均值 | 4.492 | 4.490 | 4.523 | 4.512 | 4.513 |
|  |  |
| 最小值 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 16 | 16 | 16 | 16 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 低，中，高风险洪水概率分布情况 | | | | | | |
|  | 样本数 | 平均值 | 最小值 | 25% | 75% | 最大值 |
| 低风险 | 282202 | 0.4423 | 0.285 | 0.43 | 0.46 | 0.47 |
| 中风险 | 492938 | 0.5046 | 0.475 | 0.49 | 0.52 | 0.535 |
| 高风险 | 273435 | 0.5684 | 0.54 | 0.55 | 0.58 | 0.725 |



可以看出我们分类的结果很好地满足了分类要求，接下来我们通过线性回归建立预警评价模型。

**7.3建立发生洪水不同风险的预警评价模型**

**7.3.1构建基准模型：线性回归**

线性回归是一种用于描述自变量与因变量之间线性关系的统计学方法。其基本思想是通过拟合一条直线（或超平面）来最好地拟合数据点，从而预测因变量的取值。由于该模型原理较为简单，我们选择将其作为基准模型，并借其性质来进一步探索指标数据与洪水概率的深层次关系。

下面详细介绍线性回归的过程和数学原理：

1. **线性关系建模**：线性回归假设因变量y与自变量x之间存在线性关系，可以用一条直线或超平面表示。数学表达式为

其中为自变量，为对应权重,[=()]称之为权重向量,[=()]称之为特征向量。

1. **最小化损失函数**：通常将均方误差（Mean Squared Error，MSE）作为损失函数，即：

其中为真实值，为预测值。

线性回归的目标是最小化损失函数，即通过最小化实际值与预测值之间的误差平方和来得到最优参数。采用最小二乘法得到最优参数

其中为一(n+1)]的矩阵，称之为特征矩阵（每一行均为不同样本点的特征向量的矩阵,故n+1为指标数+1，m为样本数目），[=()]是由对应样本的应变量构成的向量。

通过线性回归我们得到各个指标特征的权重如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 各指标权重（） | | | | | |
| 指标 | 季风  强度 | 地形  排水 | 河流  管理 | 森林  砍伐 | 城市化 |
| 权重 | 0.0056 | 0.0056 | 0.0057 | 0.0057 | 0.0057 |
| 指标 | 气候  变化 | 大坝  质量 | 淤积 | 农业  实践 | 侵蚀 |
| 权重 | 0.0057 | 0.0057 | 0.0056 | 0.0056 | 0.0056 |
| 指标 | 无效  防灾 | 排水  系统 | 海岸脆弱性 | 滑坡 | 流域 |
| 权重 | 0.0056 | 0.0056 | 0.0057 | 0.0056 | 0.0056 |
| 指标 | 基础设施恶化 | 人口  得分 | 湿地  损失 | 规划  不足 | 政策  因素 |
| 权重 | 0.0056 | 0.0057 | 0.0056 | 0.0056 | 0.0056 |

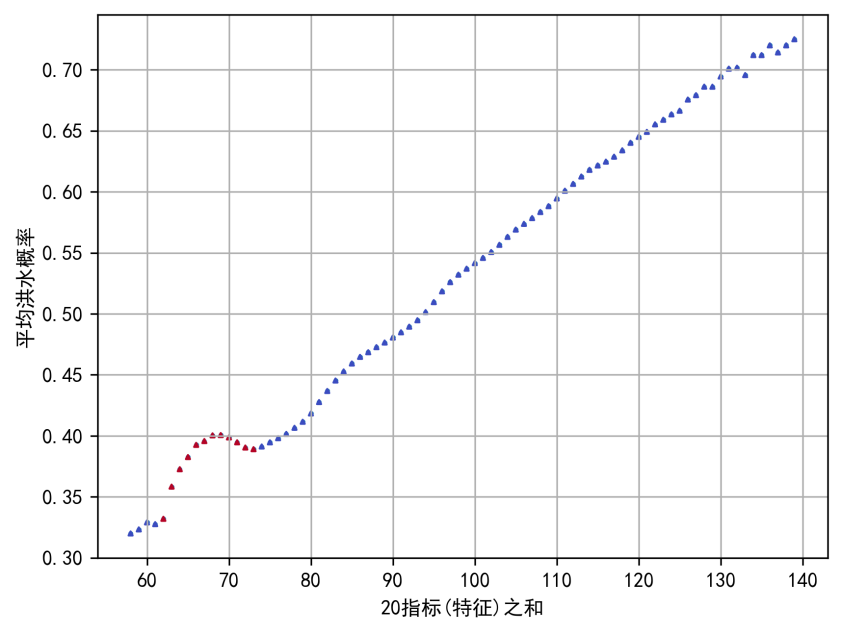
至此我们得到了洪水概率预测的基准模型，接下来我们选取关键指标用以进一步建立预警评价模型。

同时，线性回归模型得出的基准结果如下：

|  |  |
| --- | --- |
| 线性回归模型（Org-20） | |
| **R2分数** | **三分类准确率** |
| 0.84444 | 0.76743 |

**7.3.2选取关键指标**

在上述线性回归模型可以发现，指标数据未经特殊处理的情况下依旧与洪水概率有着较强的线性相关性，因此，为了构造关键引导变量，将所有指标之和求出，并与其相应洪水概率可视化如下：



可以观察到，洪水发生概率与指标和基本呈正相关。因此我们选用指标和作为关键指标，引导模型训练。同时，观察到线性关系中存在特殊值（红色部分）。为了保证模型的低灵敏度（强泛化性、强鲁棒性），将特殊值作为数据噪声放入指标考虑范围内。

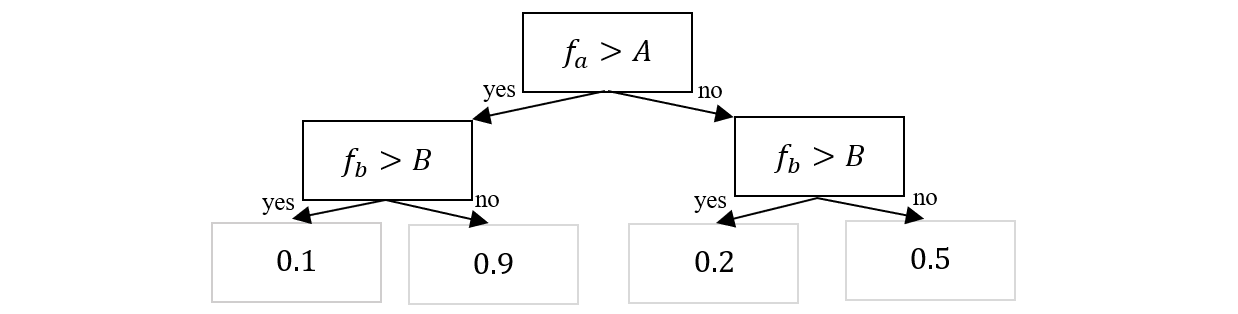
同时，由于指标维度上数据参差不齐，为了减少在特定维度上的指标数值差异，在样本维度上对所有指标进行降序排序以减少指标维度上的差异性，构造排序指标。

在这一步骤，我们通过对数据的可视化探索选取**指标和**作为一个关键指标，并且将季风强度、地形排水等指标在样本维度上从小到大排序，构造**排序指标**，作为另一个关键指标。

**7.3.3 CatBoost进一步建立模型**

为了更好地介绍CatBoost之前我们先简单地讲解几个概念：

* **决策树：**决策树是一个预测模型，它代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。如图所示就是一个决策树。



决策树的结果可以表示为：

其中是输入样本x的预测值，M为决策树的叶节点数量，是决策树第m个叶节点的区域，是第m个叶节点的类别输出，是指示函数，表示如果x属于则为1，否则为0。

* **集成学习方法：**集成学习方法是一种机器学习策略，旨在通过组合多个学习器的预测结果来改善整体的预测性能。它可以通过结合多个模型的优点，来提高泛化能力和预测准确性，特别是在处理复杂问题或数据集中存在噪声的情况下表现良好。常见的集成学习方法有Bagging（Bootstrap Aggregating）， Boosting，Stacking等等，后文使用的CatBoost就是一种通过串行训练多个弱学习器，每个学习器专注于修正前序学习器的错误，最终构建一个强大的集成模型的集成学习方法。
* **梯度提升树：**梯度提升树是一种强大的集成学习方法，通过结合多棵决策树来提升预测性能。它基于逐步优化损失函数的梯度，每一棵树都专注于拟合前一棵树的残差，最终通过组合多个模型来减少预测误差，广泛应用于回归和分类任务中，特别适合处理复杂的非线性关系和高维数据。

根据以上概念我们可以得到CatBoost的基本原理：

CatBoost是一种基于对称决策树（oblivious trees）为基学习器实现的参数较少、支持类别型变量和高准确性的GBDT框架。其过程如下：

1. **定义模型：**CatBoost回归模型可以表示为一个集成模型，由多个决策树组成。假设我们有K棵树，模型可以表示为：

其中，是输入特征向量，是第k棵树的预测函数。

1. **定义损失函数：**为了训练模型，我们需要定义一个损失函数来衡量模型的预测误差。通常在回归任务中，我们使用均方误差（MSE）作为损失函数：

其中，N是训练样本的数量，是第i个样本的真实标签。

1. **训练过程:**CatBoost使用梯度提升算法来逐步优化模型。在每一步中，根据当前模型的残差计算新的决策树，使得损失函数最小化。每棵树的建立过程可以表示为：

这里 是前k-1棵树的累积预测，L是损失函数。

1. **正则化：**CatBoost在训练过程中还包括一些正则化技术，如树的深度限制、学习率控制等，以防止过拟合并提高泛化能力。
2. **预测**：训练完成后，通过将所有树的预测累加来得到最终的预测结果：

这就是CatBoost回归模型的基本数学过程：通过梯度提升的方式，逐步构建多棵决策树，并结合它们的预测来最小化损失函数，从而得到最优的回归预测模型。

为了评估我们最后得到的回归模型，我们先引入R2分数和三分类准确率。

* **R2分数：**也称为确定系数（Coefficient of Determination），是衡量回归模型拟合优度的一个常用指标,其计算公式如下：

它直观地反映了模型预测能力的强弱，并且可以用来解释模型对数据的解释程度（值越大，说明回归效果越好）。

* **三分类准确率(ACC\_3):**三分类准确率是指在一个有三个不同类别的分类问题中，模型正确预测的样本比例。通常情况下，三分类准确率可以被定义为模型对所有三个类别中正确预测的样本数目的比例,即：

这里，“预测正确的样本数”是指模型在所有预测中正确预测的样本数目，总样本数则是所有样本的数量。显然，值越大，也说明回归效果越好。

最终，我们得到的CatBoost回归模型效果如下：

|  |  |
| --- | --- |
| CatBoost回归模型（Sorted-20，Sum-1） | |
| **R2分数** | **三分类准确率** |
| 0.86923 | 0.76290 |

**7.3.4 K折交叉灵敏性验证**

K折交叉验证（K-fold Cross-Validation）是一种常用的交叉验证技术，用于评估模型在数据集上的性能和泛化能力。它将数据集分成K个子集，每个子集称为一个折（fold）。K折交叉验证的过程如下：

1. **数据集划分：**

* 将数据集分成K个大致相等的部分（折）。
* 每个折依次作为验证集，其余K-1个折作为训练集。

1. **交叉验证过程：**

* 第一轮：将第一折作为验证集，其余K-1折作为训练集，训练模型并在第一折上进行评估。
* 第二轮：将第二折作为验证集，其余K-1折作为训练集，训练模型并在第二折上进行评估。
* 依此类推，直到第K轮。

1. **评估指标**：每轮验证后得到一个评估指标（比如准确率、精确率、召回率等）。

最终的模型性能评估通常是五轮验证结果的平均值，这样可以减少因为特定数据分割而引入的偏差。

此处我们采用**5折交叉验证**，得到的结果如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CatBoost回归模型 5折交叉验证 | | | | | | |
|  | 1折 | 2折 | 3折 | 4折 | 5折 | 均值 |
| R2分数 | 0.87019 | 0.86971 | 0.86849 | 0.86885 | 0.86892 | 0.86923 |
| ACC\_3 | 0.76247 | 0.76296 | 0.76290 | 0.76295 | 0.76324 | 0.76290 |

八、问题3模型的建立与求解

**8.1利用精选指标优化模型（基于问题1的分析）**

通过计算统计量与洪水概率的Spearman相关系数，我们得到各统计量与洪水发生概率的相关性如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 表 统计量与洪水发生概率的Spearman相关系数 | | | |
| 统计量 | 平均值 | 最小值 | 最大值 |
|  | 0.176044 | 0.170289 | 0.181399 |
| 统计量 | 25% | 50% | 75% |
|  | 0.173718 | 0.176304 | 0.178786 |

由上表可知，平均值作为统计量，与洪水发生概率的相关性居中，因此选择平均值作为划分指标来进行模型优化。因此选取出检验后Spearman相关系数高于平均的指标如下共10个：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 精选指标 | | | | |
| 基础设施恶化 | 地形排水 | 季风强度 | 大坝质量 | 河流管理 |
| 淤积 | 人口得分 | 气候变化 | 森林砍伐 | 滑坡 |

同时，加入指标和作为关键引导指标，组成全新的11维特征，训练CatBoost回归模型，性能较基准模型有所提升，而相比排序指标+指标和的模型部分伯仲，训练结果如下：

|  |  |
| --- | --- |
| CatBoost回归模型（Selected-10，Sum-1） | |
| **R2分数** | **三分类准确率** |
| 0.86923 | 0.76290 |

**8.2 五指标模型优化**

为了在仅选用5个关键指标作为特征的限制下优化问题2中所建立的CatBoost回归模型，本文通过借助划分指标来进行主成分分析，将20维指标降维至**4维关键指标**，结合**指标和**组成**5个关键指标**，实现求解得到**最优的CatBoost回归模型**。

**8.2.1主成分分析（PCA）**

* **PCA简介**

主成分分析（principal component analysis）是一种通过正交变换，将一系列可能线性相关的变量，转换成一组线性不相关的新变量的数学降维方法。而这些不相关变量也称作主成分。主成分分析在简化运算、去除噪音、数据可视化等方面起到了重要作用。

PCA算法的大致流程如下：

我们将m个n维数据记作矩阵

1. 对数据去中心化：



1. 计算协方差矩阵C：



1. 求解C的特征值矩阵（其中特征值按降序排列）,取前k列，记作矩阵:



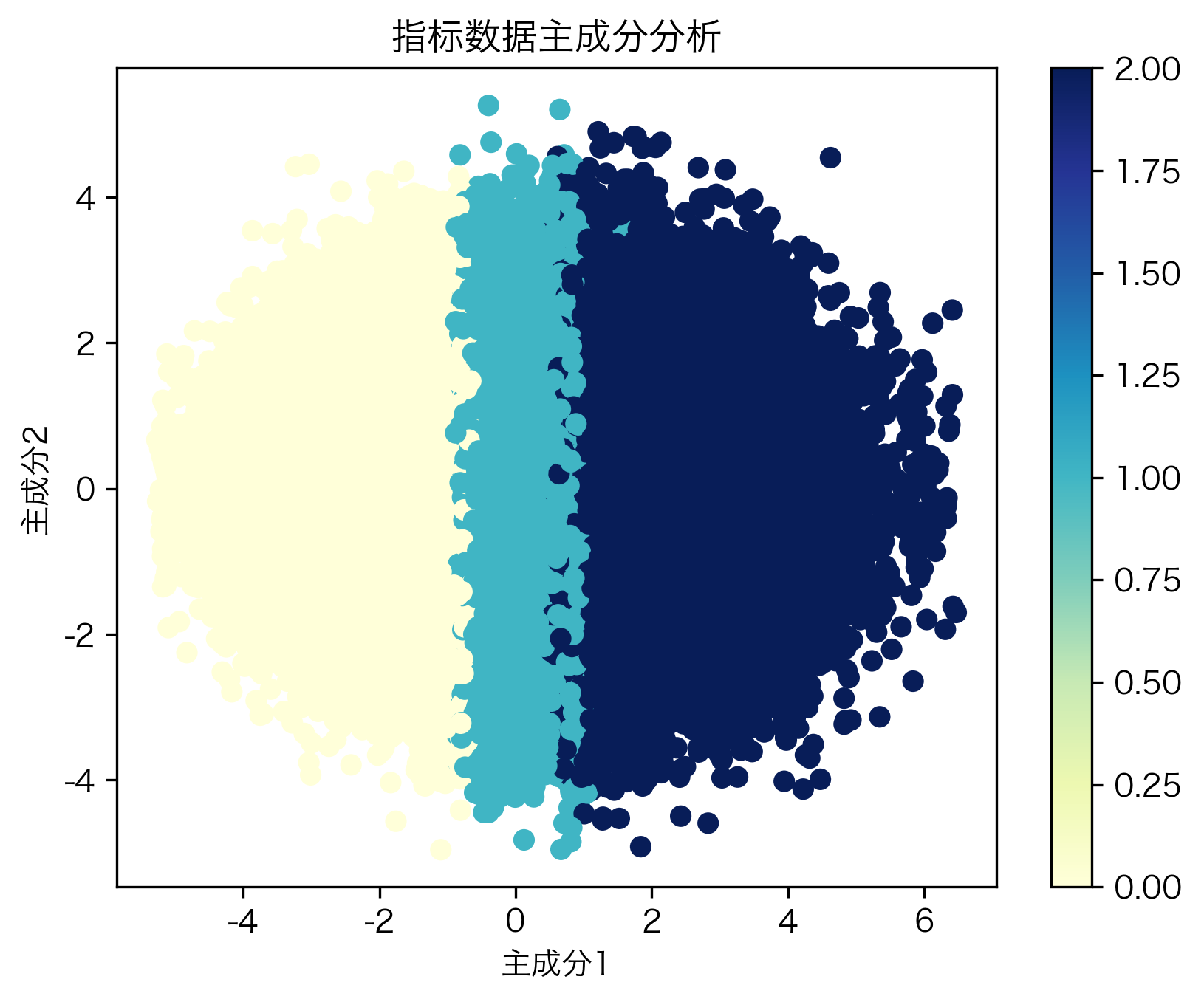
1. 将原始数据对进行投影，得到降维后数据，即：



* **应用PCA提取关键特征**

PCA降维提取关键特征的步骤如下：

1. **数据选取：**选取所有20维指标作为特征维度，洪水概率为目标维度
2. **特征标准化：**PCA对数据的尺度敏感，不同尺度的数据可能会导致主成分分析的结果不准确，因此对其进行标准化处理
3. **特征降维：**选取降维比例为20%，将20维特征降维至4维，作为关键特征
4. **PCA可视化：**展示主成分1、2的分布情况，并根据目标变量对数据点进行了颜色编码，以显示不同类别的数据分布。



**8.2.2最优CatBoost回归模型**

基于PCA降维后的4个主成分，辅以指标和组成5个关键指标，用以训练CatBoost回归模型，并实现了远超先前所有基准、优化模型的效果，实现了SOTA（State Of The Art）的最优模型：

|  |  |
| --- | --- |
| SOTA CatBoost回归模型（PCA-4，Sum-1） | |
| **R2分数** | **三分类准确率** |
| **0.97369** | **0.90594** |

具体模型优化效果及其与基准模型的对比如下表所示：

|  |  |  |
| --- | --- | --- |
| 洪水概率预测回归模型 | | |
| **模型** | **R2分数** | **ACC\_3** |
| 线性回归(Org-20) | 0.84444 | 0.76743 |
| XGBoost(Org-20) | 0.80975 | 0.74846 |
| CatBoost(Org-20) | 0.84634 | 0.76736 |
| CatBoost(Sorted-20,Sum-1) | 0.86923 | 0.76290 |
| CatBoost(Selected-10,Sum-1) | 0.86691 | 0.76281 |
| SOTA **CatBoost(PCA-4,Sum-1)** | **0.97369** | **0.90594** |

由此可清晰看到本文对模型优化的进行过程及特征工程带来的极佳优化效果。

九、问题4模型的建立与求解

基于问题2选取的回归模型CatBoost，加以问题3对指标进行的关键指标选取，实现了最优的回归模型，并该模型与特征工程应用于test.csv数据，并将预测结果填写进入submit.csv中。

按照问题1中的数据分析结果，测试集的频数分布直方图与训练集基本相同，而训练集的洪水概率的分布情况基本符合正态分布，**因此可以假设测试集的预测洪水概率分布情况也应基本尽量符合正态分布**。

对预测结果的概率密度直方图可视化如下：

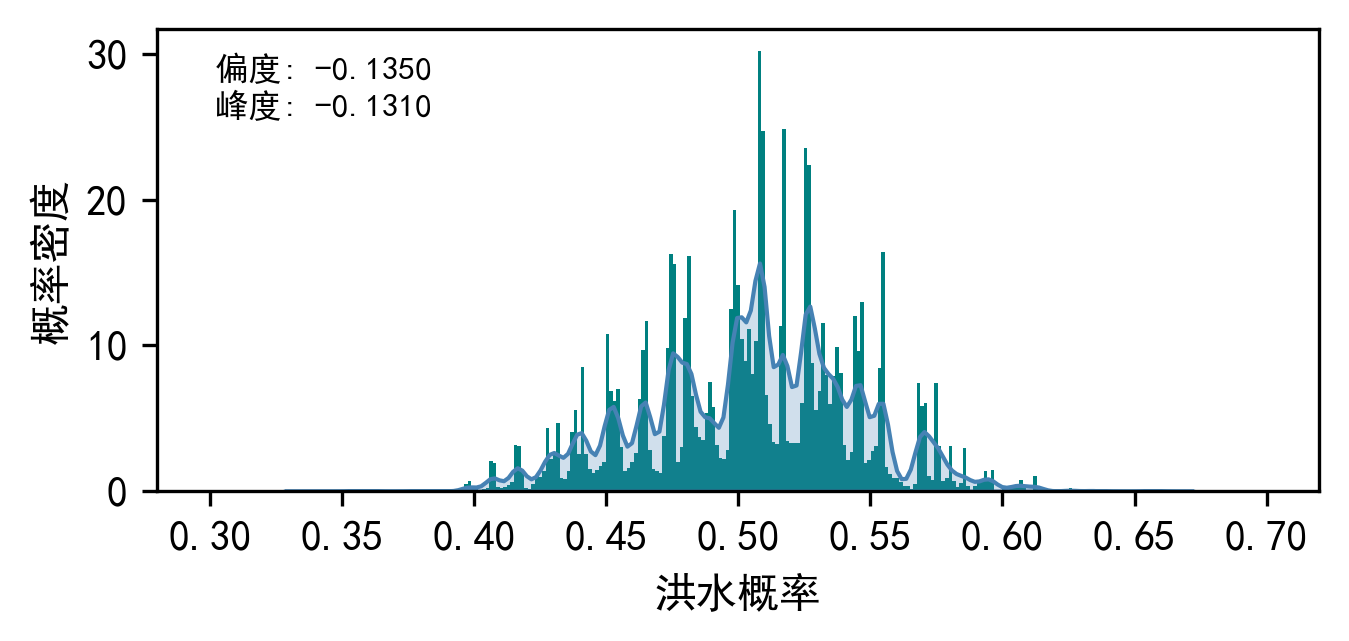
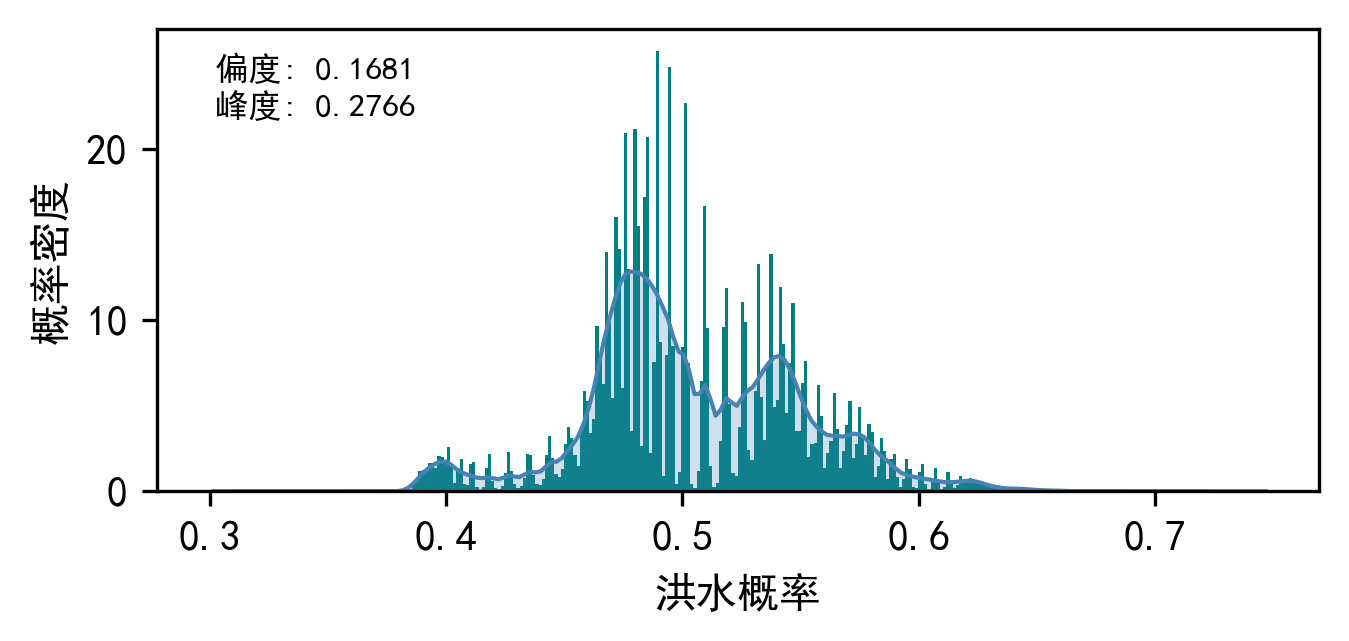
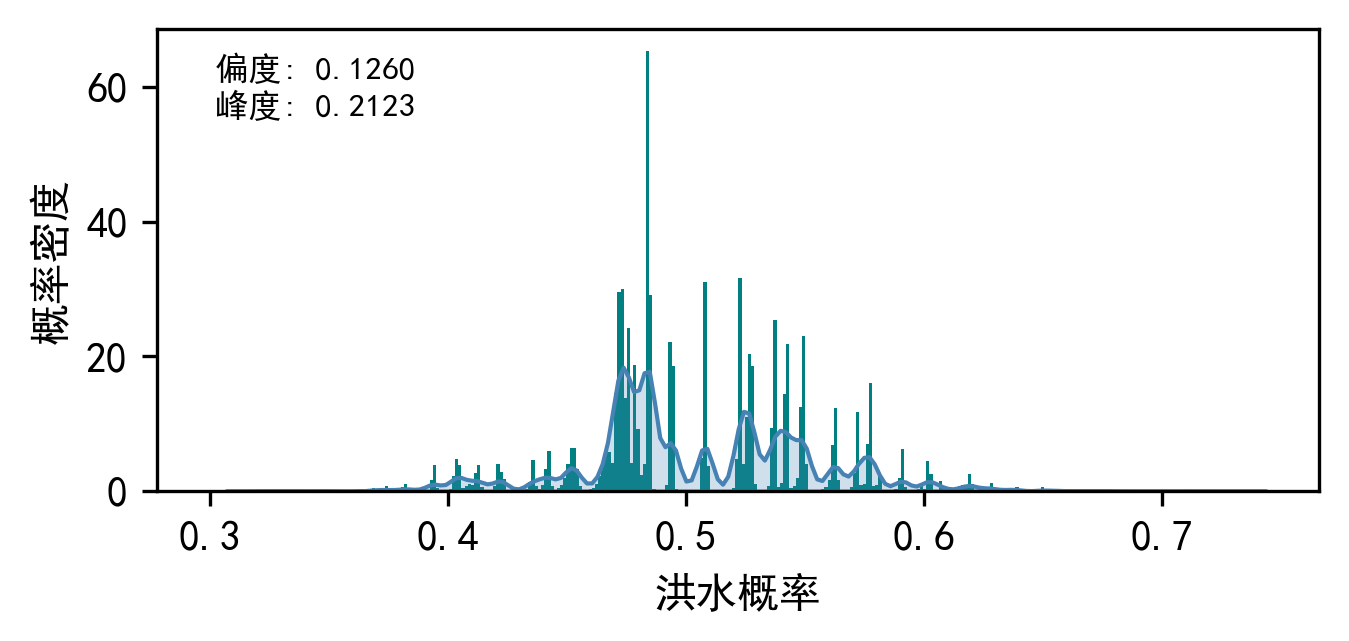


图 洪水发生概率的统计直方图

该预测结果分布情况没有明显的偏向性，基本符合正态分布，对比其他性能较弱的两个CatBoost回归模型的预测结果概率分布直方图如下：



性能较弱的两个模型有着明显的预测偏向性，并不符合正态分布，可见，性能最优预测模型的预测结果是较为可信的，这也验证了该模型的鲁棒性。

对最优模型预测结果按正态分布随机采样1000条数据，可视化为折线图如下：

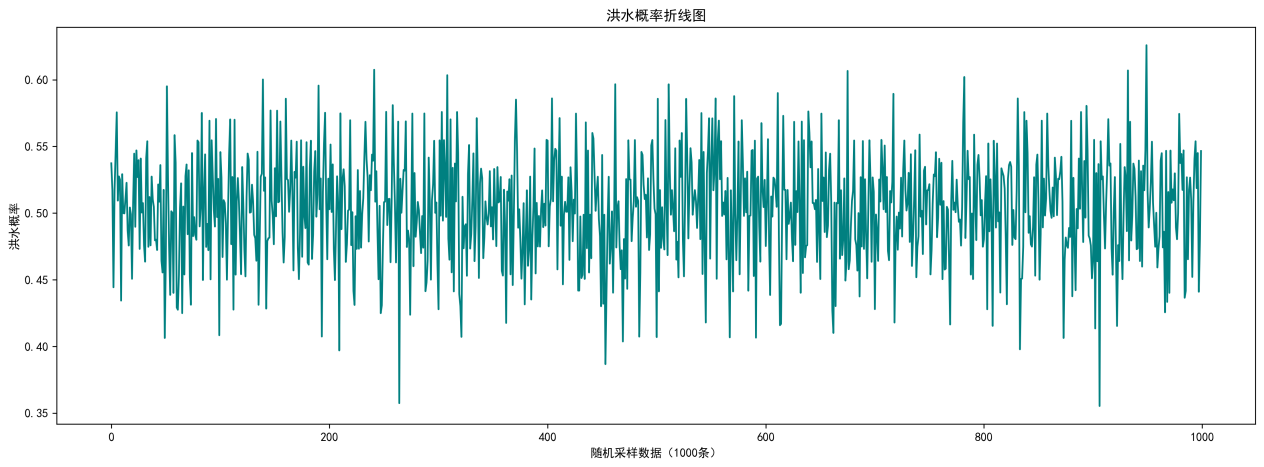


图 洪水发生概率的统计折线图

可见数据基本在0.5周围上下浮动，异常值较少，并不影响整体数据分布情况。

十、模型的评价与推广

十一、参考文献

十二、附录

**参考文献 （可另起一页）**

参考文献的编号，如[1][3]等；引用书籍还必须指出页码。参考文献按正文中的引用次序列出，其中：**书籍的表述方式为**

[编号] 作者，书名，出版地：出版社，出版年。

**参考文献中期刊杂志论文的表述方式为**

[编号] 作者，论文名，杂志名，卷期号：起止页码，出版年。

**参考文献中网上资源的表述方式为**

[编号] 作者，资源标题，网址，访问时间（年月日）。

**附录（另起一页）**