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		Apmcm24102513

基于机器学习的洪水灾害预测与预防

摘要

本文使用了基于皮尔逊相关系数的相关分析模型、K-Means 聚类模型、主成分回归模型、XGBoost 预测模型等机器学习模型，以及 K-S 检验和 Q-Q 图等统计分析的方法，对洪水灾害预测和预防进行了深入研究。

针对问题一，首先，我们对 train.csv 的数据进行预处理，确认其中没有缺失值与异常值，其次，可视化 20 个指标的概率分布，得到各指标基本符合正态分布的结论，然后，使用皮尔逊相关分析的方法计算出各指标与洪水发生概率的相关性，得到基础设施恶化与洪水爆发相关性最高，海岸脆弱性与洪水爆发相关性最低等结论。最后，对相关性计算结果进行解释，并提供对老旧的防洪设施进行维护和升级等建议。

针对问题二，首先，我们使用 K-means 聚类分析将洪水概率分成三类，其中中等风险数量最多，有 455573 组数据。其次，通过可视化比较各指标在不同类别下的平均值，得到每一个指标平均值都是高风险大于中风险大于低风险。然后，建立基于主成分回归分析的风险预警评价模型，将 20 个特征指标降维成 5 个主成分，其中主成分 2 的权重最高，达到了 0.224，最后，通过对输入进行扰动，观察模型性能变化情况，得到我们的回归模型对输入小幅度波动具有一定鲁棒性。

针对问题三，首先，我们根据皮尔逊相关系数得分，对 20 个指标进行排序，在规定指标个数为 5 的情况下，我们选择 5 个的相关性最大的指标作为特征指标。其次，将筛选后的指标数据通过 XGBoost 预测模型进行训练，将百分之 70 作为训练集，百分之 30 作为测试集，得到训练集和测试集的 RMSE 都收敛到 0.046 的预测模型。

针对问题四，首先，我们将 test.csv 的相关数据带入问题二建立的基于主成分分析回归模型中，将预测的结果通过概率直方图和折线图进行可视化分析，得到数据具有正态性和集中趋势，并进一步通过 K-S 检验和 Q-Q 图验证，得到样本数据基本符合正态分布的结果。

关键词：洪水灾害预测，相关分析，K-means 聚类，主成分回归，XGBoost，MATLAB

目录

一、 问题重述

1.1 问题背景

洪水作为一种自然灾害，其频发和严重程度随着气候变化和人类活动的加剧而不断增加。2023 年全球多地发生洪水灾害，造成了巨大的经济损失和人员伤亡。因此，对洪水的监测、预测和预防显得尤为重要。本次竞赛的题目聚焦于利用大规模数据分析和建模技术，预测洪水发生的概率，旨在提升对洪水灾害的应对能力。

1.2 题设数据

附件 train.csv：包含超过 100 万条洪水事件数据，每条数据记录了洪水事件的 ID、季风强度、地形排水、河流管理、森林砍伐、城市化、气候变化、大坝质量、淤积、农业实践、侵蚀、无效防灾、排水系统、海岸脆弱性、滑坡、流域、基础设施恶化、人口得分、湿地损失、规划不足、政策因素和发生洪水的概率。

附件 test.csv：包含超过 70 万条洪水事件数据，这些数据记录了洪水事件的 ID 和上述 20 个指标的得分，但没有记录洪水发生的概率。

附件 submit.csv：包括 test.csv 中的洪水事件 ID，需要填入预测的洪水发生概率。

1.3 需解决的问题

问题一：分析附件 train.csv 中的数据，确定并可视化上述 20 个指标中哪些与洪水发生有显著关联，哪些关联性较低，并分析可能原因。基于分析结果，提出合理的预防洪水灾害的建议和措施。

问题二：将附件 train.csv 中记录的洪水发生概率进行分类，识别高、中、低风险的洪水事件，并分析各风险类别的指标特征。选取合适的指标，计算其权重，建立洪水风险预警模型，并对模型进行灵敏度分析。

问题三：基于问题一中对指标的分析结果，建立洪水发生概率的预测模型，选取合适的指标来预测洪水发生的概率，并验证模型的准确性。探讨如何在仅使用 5 个关键指标的情况下，调整并改进预测模型。

问题四：利用问题二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，并将预测结果填入附件 submit.csv 中。绘制洪水发生概率的直方图和折线图，分析概率分布是否服从正态分布。

二、 问题分析

2.1 问题一的分析

问题一需要分析并可视化 20 个指标，找出与洪水发生密切相关和相关性不大的指标，分析可能原因并针对洪水的提前预防提出合理的建议和措施。首先，我们对 train.csv 的数据进行预处理，观察其中是否有缺失值和异常值，其次，对 20 个指标的概率分布可视化，根据概率分布情况，选择合适的相关分析方法，然后，根据相关分析计算出的相关性，选出与洪水发生密切相关和相关性不大的指标，最后，对结果进行解释，并对预防洪水提供一些建议和措施。

2.2 问题二的分析

问题二要求我们将洪水概率聚类成低中高三种类别，分析不同类别的指标特征，然后选取合适指标建立风险预警评价模型。首先，我们使用 K-means 聚类分析将洪水概率分成三类，通过可视化比较各特征在不同类别下的平均值，得到不同类别的指标特征，其次，建立基于主成分回归分析的风险预警评价模型，将 20 个特征指标降维成主成分，得到各主成分的权重，并对模型的性能进行评价，最后，通过对输入进行扰动，对风险预警评价模型的进行灵敏度分析。

2.3 问题三的分析

问题三要求我们基于问题一中指标分析结果，从 20 个指标中选取合适指标预测洪水发生概率，并验证模型的准确性，给出在仅用五个关键指标的情况下，模型应该如何调整。首先，我们根据问题一皮尔逊相关分析得出的各指标与洪水发生相关性，对 20 个指标进行排序，在规定指标个数的情况下，我们从上而下选择对应个数的指标。其次，我们将筛选后的指标数据通过 XGBoost 预测模型进行训练，将百分之 70 作为训练集，百分之 30 作为测试集，通过绘制训练误差和测试误差随迭代次数的变化情况，对模型训练效果进行评价。

2.4 问题四的分析

问题四要求使用问题 2 中的预测模型对 test.csv 中的洪水发生概率进行预测，并分析结果是否满足正态分布。首先，我们将 test.csv 的相关数据带入问题二中基于主成分分析回归模型中，将预测的结果通过概率直方图和折线图进行可视化分析，并进一步通过 K-S 检验和 Q-Q 图验证预测结果是否符合正态分布。

三、 模型假设

假设一：所用的数据都是可信的，不存在编造数据的情况。

假设二：环境与政策在模型预测期间不发生重大改变。

假设三：每个地区对文本类型的变量进行编码时所用的方法都是一样的。

假设四：洪水灾害的发生机制可以通过机器学习方法模拟。

四、 符号说明

符号	符号说明
X	样本数据的特征变量
Y	洪水发生概率
Z	主成分分析后的特征变量
n	样本量
r	皮尔逊相关系数
F(x)	理论分布函数
Fn(x)	经验分布函数
Dn	K-S 统计量
p	显著性水平下的 p 值
μ	均值
σ	标准差
X_i	第 i 个样本的特征变量值
Y_i	第 i 个样本的实际值
b	回归系数
a	截距
e	误差项
R^2	决定系数
MSE	均方误差
RMSE	均方根误差
k	聚类中心数
c	信号传播速度

五、 问题一模型的建立与求解

5.1 数据处理

在进行数据分析前，我们需要对数据进行处理，剔除其中的异常值，我们先对 train.csv 中的数据进行数量统计，得到结果如下表。

表 1 train.csv 数据统计

时间	个数	标准差
地形排水	1048575	2.056183
基础设施恶	1048575	2.093638
大坝质量	1048575	2.071778
河流管理	1048575	2.051735
季风强度	1048575	2.08314
人口得分	1048575	2.05792
淤积	1048575	2.082146
气候变化	1048575	2.065527
森林砍伐	1048575	2.067986
滑坡	1048575	2.083124
无效防灾	1048575	2.078136
政策因素	1048575	2.072499
湿地损失	1048575	2.089066
农业实践	1048575	2.077774
规划不足	1048575	2.082354
流域	1048575	2.064532
城市化	1048575	2.074506
海岸脆弱性	1048575	2.068883
侵蚀	1048575	2.081685
排水系统	1048575	2.090317
洪水概率	1048575	0.051031

从描述性统计结果来看，数据中无缺失值，各特征的标准差相对一致，洪水概率由于数值位于 0-1 之间，较为集中，因此标准差非常低。综合分析后，我们认为数据中无异常值。

5.2 可视化概率分布

为了选择合适的相关分析方法，我们需要对样本数据的概率分布进行分析，

如果样本近似于正态分布，我们可以使用皮尔逊相关系数进行相关分析，通过MATLAB 软件对附件进行概率分布可视化分析，得到 20 个特征值的概率分布如下。

图 1 20 个特征值的概率分布情况

从上图可知，各特征值的概率分布基本满足正态分布，我们可以使用皮尔逊相关分析进行求解。

5.3 基于皮尔逊相关系数的相关分析模型

皮尔逊相关系数是描述 2 个定距变量间联系紧密程度，衡量变量 X 和 Y 之间的线性相关关系的参数，其值介于 -1 与 1 之间，一般用 r 表示，计算公式为

$$r_{xy}=\frac{n\sum XY-\sum X\sum Y}{\sqrt{\left[N\sum X^2-\left(\sum X\right)^2\right]\left[N\sum Y^2-\left(\sum Y\right)^2\right]}} \tag{1}$$

式中，n 为样本量，X、Y 分别为 2 个变量的值。

若 r>0，反映 2 个变量为正相关关系;若 r=0，则表示 2 个变量不相关;若 r<0，反映 2 个变量为负相关关系。r 的绝对值越大，表明相关性越强。

通常情况下，按如下取值区间对相关性强弱进行判断，r 的绝对值在 0.8~1.0 之间为极强相关，0.6~0.8 之间为强相关，0.4~0.6 之间为中等程.度相关，0.2~0.4 之间为弱相关，0.0~0.2 之间极弱相关或无相关。

该方法运算效率高，根据原理能够快速建立数学计算模型，进行客观的定量分析，对各项参数定量计算，避免了定性分析的不确定性，且实用性强。

（1）标准化指标

不同变量的量纲之间存在差异，为了统一量纲，我们将数据按如下步骤进行标准化

- 先计算出各指标变量的算术平均值和标准差:

算术平均值：

$$\overline{X_i} = \frac{\sum_{j=1}^{n} X_{ij}}{4} \tag{2}$$

标准差：

$$S_i = \sqrt{\frac{\sum_{j=1}^{n} \left(x_{ij} - \overline{X_i} \right)}{n-1}} \tag{3}$$

- 将指标进行标准化处理

$$Z_{ij} = \frac{X_{ij} - \overline{X_i}}{S_i} \tag{4}$$

Zij 为第 i 个学生的第 j 个指标经标准化后的值

（2）模型求解

将处理过后的数据在 Matlab 软件上运行，得到各特征与洪水爆发的相关性如下图所示。

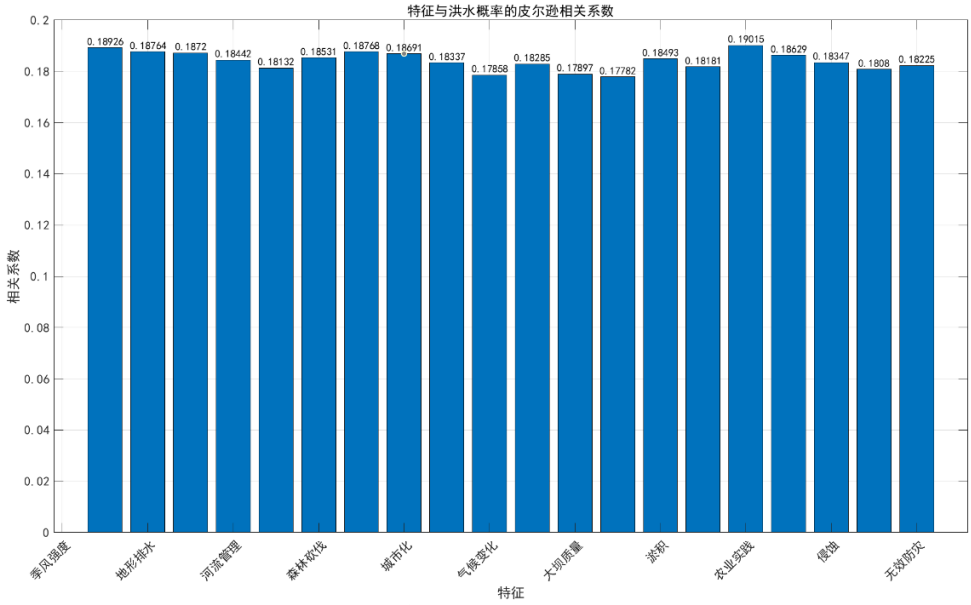

图 2 各特征与洪水爆发的相关性

从上图可知，基础设施恶化（相关系数：0.19015）、季风强度（相关系数：0.18926）、大坝质量（相关系数：0.18768）、地形排水（相关系数：0.18764）和河流管理（相关系数：0.18720）与洪水发生密切相关，而侵蚀（相关系数：0.17858）、排水系统（相关系数：0.17897）、海岸脆弱性（相关系数：0.17782）

与洪水发生相关性较低。

5.4 结果分析与防洪建议

我们对各指标对洪水发生的影响进行分析，得到结论如下：

第一，基础设施恶化可能导致防洪设施失效、堤坝崩塌和排水系统堵塞，从而增加洪水风险。第二，强季风带来大量降水，直接增加洪水发生的可能性。第三，大坝质量差可能导致在洪水期间大坝崩溃，增加下游地区的洪水风险。第四，良好的地形排水和河流管理可以有效减轻洪水的影响，因此这些因素与洪水发生密切相关。

根据我们的分析结果，我们给出以下建议与措施：

1. 提升基础设施质量：对老旧的防洪设施进行维护和升级，确保它们在洪水期间能够正常工作。建立并执行严格的基础设施质量标准，以减少因基础设施失效导致的洪水风险。
2. 加强季风监测与预警：建立完善的气象监测系统，实时监测季风强度和降雨量，及时发布洪水预警信息。
3. 提高大坝质量和管理水平：定期检查和维护大坝，及时修复发现的隐患。建立应急预案，在大坝出现险情时迅速采取措施，减少洪水风险。

六、 问题二模型的建立与求解

6.1 基于 K-means 的聚类分析模型

6.1.1 K-means 分析的原理

K-Means 是一种常用的非监督学习算法，用于将数据集分成 k 个不同的聚类。其目标是将数据集中的数据点划分到 k 个聚类中，使得同一聚类中的数据点之间的相似度最大，不同聚类之间的相似度最小。K-Means 聚类分析的步骤如下：

1. 选择初始聚类中心：随机选择 k 个数据点作为初始聚类中心。
2. 分配数据点：根据每个数据点与各聚类中心的距离，将数据点分配到距离最近的聚类中心。
3. 更新聚类中心：计算每个聚类的质心，将质心作为新的聚类中心。

4. 重复步骤 2 和 3：直到聚类中心不再发生显著变化，或者达到预定的迭代次数。

在上述步骤中，本文使用的距离测量方法是欧氏距离：

$$\text{distance}(x,c)=\sqrt{\sum_{i=1}^{n}(x_i-c_i)^2} \tag{5}$$

其中，x 表示数据点，c 表示聚类中心，n 表示数据的维度。

6.1.2 聚类结果与特征分析

将 train.csv 中洪水发生概率数据带入聚类分析模型中，分成三类，得到各类别的聚类数量和聚类中心如下表。

表 2 各类别的聚类数量和聚类中心

聚类类别	聚类中心	聚类数量
0	0.497445447	455573
1	0.560868807	346148
2	0.438348187	246854

从上表可知，高中低风险分别对应类别 1、0、2，聚类中心为 0.561，0.497 和 0.438，聚类数量为 346148、455573 和 246854。

为了更好的体现各类别的特征特性，我们将三种类别对应的 20 个特征平均值进行比较分析，得到特征对比结果图如下。

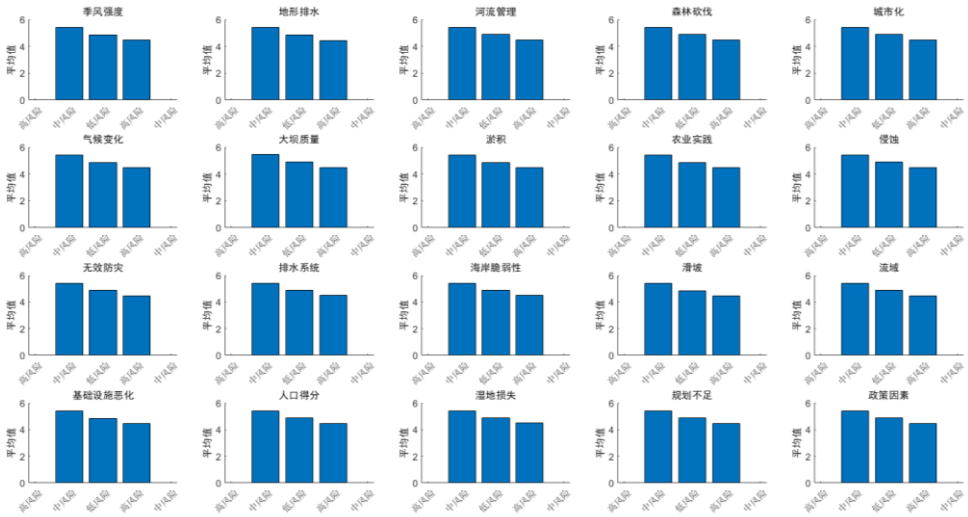

图 3 不同类别的特征对比结果图

从上图可知，每一个特征都是高风险大于中风险大于低风险，规律十分明显，但是为了降低计算维度，我们需要选取重点指标进行分析，根据我们的聚类

情况，我们认为各特征与洪水发生概率有明显的线性关系。

因此，我们选择使用主成分回归分析的方法进行关键指标的选取和各关键指标在风险评价模型的权重评估。

6.2 基于主成分回归分析的风险的预警评价模型

6.2.1 主成分回归分析的原理

主成分回归（Principal Component Regression，简称 PCR）是一种结合主成分分析（PCA）和线性回归的方法，适用于处理多重共线性问题，并简化模型。其基本思想是通过主成分分析将原始高维数据降维，提取出少数几个主要成分，再利用这些主要成分进行线性回归，从而提高模型的稳定性和预测精度。

具体来说，主成分回归分析包括以下步骤：

1. 数据标准化：首先对数据进行标准化处理，使每个特征的均值为 0，标准差为 1，消除量纲差异的影响。
2. 主成分分析（PCA）：对标准化后的数据进行主成分分析，计算出特征矩阵的主成分（principal components）。这些主成分是原始特征的线性组合，且相互正交。主成分的选择通常依据解释总方差的比例，选取能够解释大部分方差的前几个主成分。
3. 线性回归：利用选取的主成分作为新的特征变量，进行线性回归分析，构建预测模型。
4. 模型评估：通过计算决定系数（R^2）、均方误差（MSE）等指标评估模型性能。

6.2.2 关键特征选取和模型性能评价

将标准化处理后的 train.csv 数据带入主成分回归分析中，得到各主成分的计算公式如下表。

表 3 主成分计算公式

特征	PC1	PC2	PC3	PC4	PC5
季风强度	0.073	-0.075	0.058	-0.120	-0.196
地形排水	0.130	0.234	0.282	-0.057	-0.177
河流管理	0.020	0.357	-0.269	0.185	-0.049
森林砍伐	0.189	0.251	-0.029	-0.333	0.030
城市化	-0.247	-0.162	0.344	0.222	-0.075
气候变化	0.005	0.379	0.017	-0.223	0.048
大坝质量	0.069	0.140	0.376	0.063	-0.007

特征	PC1	PC2	PC3	PC4	PC5
淤积	-0.120	-0.211	-0.172	0.076	0.095
农业实践	0.070	-0.029	0.205	0.225	0.054
侵蚀	-0.495	-0.280	0.172	-0.072	0.036
无效防灾	0.283	-0.311	-0.085	-0.108	-0.015
排水系统	-0.049	0.106	0.245	-0.419	0.056
海岸脆弱性	0.588	-0.231	-0.131	0.236	-0.221
滑坡	-0.141	0.174	-0.210	-0.086	0.227
流域	0.216	-0.070	0.197	0.254	0.520
基础设施恶化	-0.034	0.255	-0.284	0.061	0.002
人口得分	-0.232	0.182	-0.143	0.457	0.145
湿地损失	-0.101	-0.137	0.018	-0.117	-0.469
规划不足	-0.226	-0.026	-0.320	0.093	-0.400
政策因素	-0.019	-0.349	-0.334	-0.349	0.363

各主成分的权重及主成分回归公式为：

$$y = -0.193Z_1 + 0.224Z_2 - 0.061Z_3 + 0.042Z_4 + 0.134Z_5 + \varepsilon \quad (5)$$

第二个主成分权重最大，达到了 0.224。对主成分回归模型进行性能分析，得到结果如下表：

表 4 主成分回归模型性能分析

评价参数	MSE	RMSE	R^2
数值	0.043	0.051	0.42

从性能评价结果的 MSE 为 0.043 可知，模型在预测洪水发生概率时具有相对较高的精度，从 R 方为 0.42 可知，我们建立的回归模型有一定的预测能力。

6.3 灵敏度分析

为了研究我们模型对输入变量变化的响应情况，我们通过对各数据的一个或多个随机指标进行扰动，观察模型最终的性能分析变化情况，得到灵敏度分析结果如下图所示。

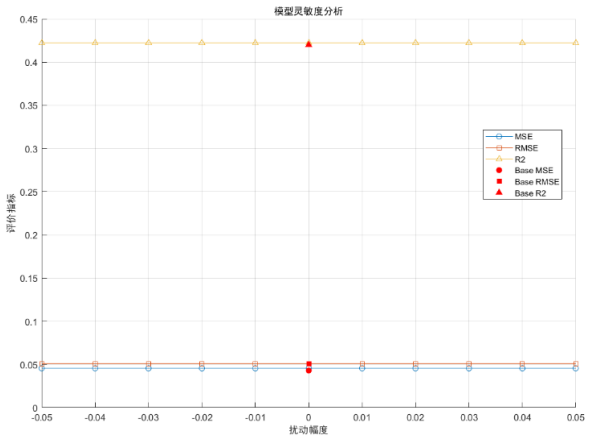

图 4 灵敏度分析

从上图可知，我们的模型最终结果对输入的扰动基本不发生改变，有一定的稳定性。

七、 问题三模型的建立和求解

7.1 基于问题一相关分析结果的特征选取

基于问题一的相关分析模型，我们可以对各特征与洪水爆发相关性进行排序，排序结果如下表。

表 5 问题一的相关系数排序

排序	特征	相关系数	排序	特征	相关系数
1	季风强度	0.189	11	无效防灾	0.183
2	地形排水	0.188	12	排水系统	0.179
3	河流管理	0.187	13	海岸脆弱性	0.178
4	森林砍伐	0.184	14	滑坡	0.185
5	城市化	0.181	15	流域	0.182
6	气候变化	0.185	16	基础设施恶化	0.190
7	大坝质量	0.188	17	人口得分	0.186
8	淤积	0.187	18	湿地损失	0.183
9	农业实践	0.183	19	规划不足	0.181
10	侵蚀	0.179	20	政策因素	0.182

在只能选取五个特征的情况下，我们选取的五个特征即为相关性排名前五的特征。

通过问题一的相关分析模型，我们得到排名前五的特征分别为：基础设施恶化（相关系数：0.19015）、季风强度（相关系数：0.18926）、大坝质量（相关系数：0.18768）、地形排水（相关系数：0.18764）和河流管理（相关系数：0.18720）。我们将这五个指标分别用 X1-X5 表示。

7.2 基于改进的 XGBoost 预测模型

7.2.1 XGBoost 的基本原理

XGBoost（Extreme Gradient Boosting）是由 Tianqi Chen 开发的一种高效且灵活的梯度提升（Gradient Boosting）框架。它因其高性能和高准确性在机器学习竞赛和实际应用中广泛使用。下面是对 XGBoost 原理的详细解释。

XGBoost 是一种集成学习方法，集成了多个弱学习器（通常是决策树），通过逐步减少预测误差来提高模型的性能。它基于梯度提升框架，主要思想是通过不断添加新模型来修正之前模型的预测误差。具体步骤如下：

- 初始化模型 $f_0(x)$，本文使用随机常数值初始化。
- 对于每一轮迭代 t：计算当前模型的残差，即真实值与当前模型预测值之间的差异。使用残差作为目标变量，训练一个新的弱学习器 $h_t(x)$。将新的弱学习器加入模型，其中η为学习率，更新模型：

$$f_t(x) = f_{t-1}(x) + \eta \cdot h_t(x) \tag{6}$$

- 最终得到的模型是所有弱学习器的加权和。

7.2.2 XGBoost 的改进

XGBoost 在传统梯度提升的基础上进行了以下几项的改进，以提高性能和处理能力：

第一是正则化，XGBoost 在损失函数中添加了正则化项，以控制模型的复杂度，防止过拟合。正则化项包括树的叶子数和叶子节点权重的 L1 和 L2 正则化。

第二是二阶导数信息，XGBoost 不仅使用了一阶导数（梯度），还使用了二阶导数（Hessian 矩阵）来更新模型。这使得模型训练更加稳定和高效。

第三是分裂节点的优化，XGBoost 使用贪婪算法和近似算法相结合的方法来寻找最佳分裂点，提高了树分裂的效率。

第四是树的并行构建，XGBoost 通过并行计算构建树的分裂点，大幅提高了训练速度。

第五是剪枝，在构建树时，XGBoost 通过“最大深度”参数和“最小分裂损失”参数来控制树的深度和复杂度。

第六是行采样和列采样，XGBoost 在训练过程中对训练数据进行行采样和列采样，这不仅减少了计算量，还提高了模型的泛化能力。

7.2.3 相关参数设定

XGBoost 可以应用于多种任务，包括回归、分类和排序。针对不同的任务，XGBoost 定义了不同的损失函数，本问属于回归问题，定义的损失函数为：

$$L(y, \bar{y}) = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \tag{7}$$

XGBoost 还有许多能够影响训练效果的超参数，具体参数及在本问中的设定值如下表所示。

表 6 XGBoost 参数设置

参数名称	作用	设定值
max_depth	树的最大深度，防止过拟合	6
min_child_weight	叶子节点最小权重，控制叶子节点的样本权重和	1
gamma	分裂节点的最小损失减少，控制节点分裂	0.1
subsample	样本采样率，防止过拟合	0.8
colsample_bytree	特征采样率，防止过拟合	0.8
lambda	L2 正则化项权重	1
alpha	L1 正则化项权重	0
eta	学习率，控制每次更新的步长	0.1

7.2.4 对训练结果进行评价

为了直观的观察我们模型的训练过程，我们将训练误差和测试误差随迭代次数的变化进行可视化，得到训练可视化图如下。

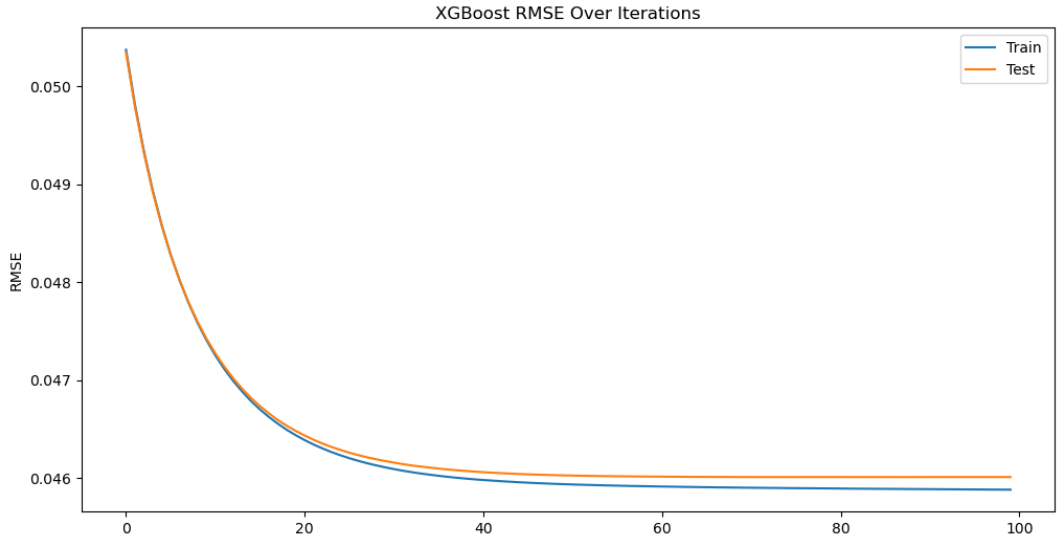

图 5 XGBoost 算法的训练可视化图

从上图可知，在初始迭代次数较低时，训练和测试集的 RMSE 快速下降。表明模型在初期通过不断学习样本数据，迅速提升了拟合效果，减少了误差。随着迭代次数增加，训练集和测试集的 RMSE 逐渐趋于平稳。表明模型逐渐接近收敛状态，进一步的迭代带来的误差减少非常有限。

整个训练过程中，训练集和测试集的 RMSE 曲线非常接近，且没有明显的发散。这表明模型在训练过程中没有发生明显的过拟合现象，模型的泛化性能较好。在迭代 100 次后，训练集的 RMSE 略低于测试集的 RMSE，但差距非常小。训练集和测试集的 RMSE 都收敛到约 0.046 的水平，这表明模型对训练数据和未见过的测试数据都有较好的预测能力。

八、 问题四模型的建立与求解

8.1 基于主成分回归分析模型的洪水概率预测

在问题二中，我们将 20 个与洪水相关的特征降维成了五个主成分，主成分与各特征的关系见表 x，并且通过训练数据得到了各主成分的权重，接下来，我们使用训练好的预测模型对 test.csv 中的数据集进行洪水概率预测。

将预测的结果填入 submit.csv 中，部分预测结果如下表所示。

表 7 洪水概率的部分预测结果

id	洪水概率预测	id	洪水概率预测
1117957	0.522109	1117974	0.51807
1117958	0.48093	1117975	0.556477
1117959	0.480923	1117976	0.490313
1117960	0.486545	1117977	0.519714
1117961	0.508538	1117978	0.53393
1117962	0.524489	1117979	0.562703
1117963	0.524489	1117980	0.544067
1117964	0.544715	1117981	0.526182
1117965	0.504581	1117982	0.500395
1117966	0.573883	1117983	0.527233
1117967	0.508514	1117984	0.528926
1117968	0.470923	1117985	0.558881
1117969	0.478097	1117986	0.522763
1117970	0.535407	1117987	0.530568
1117971	0.515048	1117988	0.534782
1117972	0.493175	1117989	0.503423
1117973	0.53956	1117990	0.501315
…	…	…	…
1118050	0.514268222	1118055	0.531227162
1118051	0.468119183	1118056	0.524049376
1118052	0.51962241	1118057	0.524385687
1118053	0.521613139	1118058	0.545353144
1118054	0.544430941	1118059	0.469132166
…	…	…	…

为了判断所得结果是否符合正态分布，我们可以通过洪水概率直方图和折线图进行初步判断，如果初步看出具有正态分布的趋势，可以使用 K-S 检验，Q-Q 图对概率分布情况进行进一步分析。

8.2 概率直方图和折线图

8.2.1 直方图分析

直方图用于展示数据的频率分布，通过观察直方图可以了解数据的分布形态。将 submit.csv 中的洪水概率结果进行直方图分析，得到概率直方图如下。

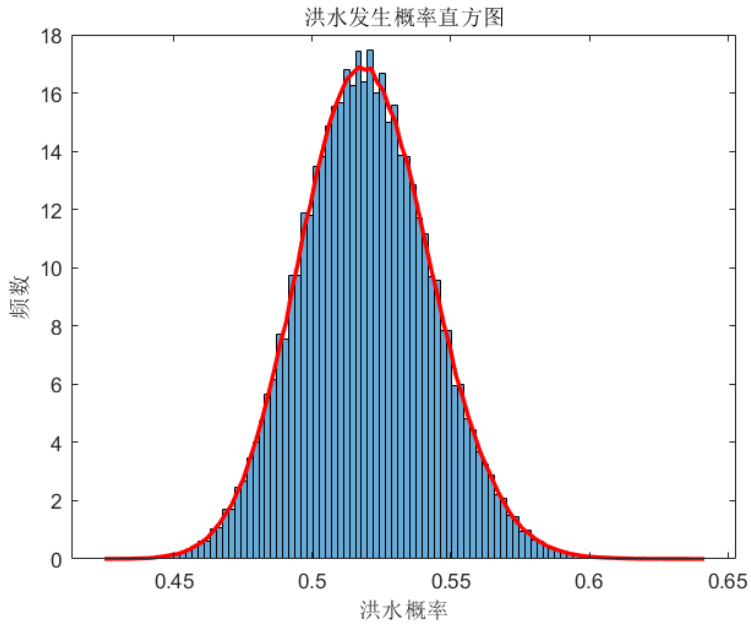

图 6 洪水概率直方图分析

从上图可知，洪水发生概率的数据呈现钟形分布，较为对称，接近正态分布，大多数事件的洪水发生概率集中在 0.5 左右。

8.2.2 折线图分析

折线图展示数据的排序趋势，通过观察折线图可以了解数据的变化趋势，我们对 submit.csv 中的洪水概率结果进行折线图分析，得到折线图如下。

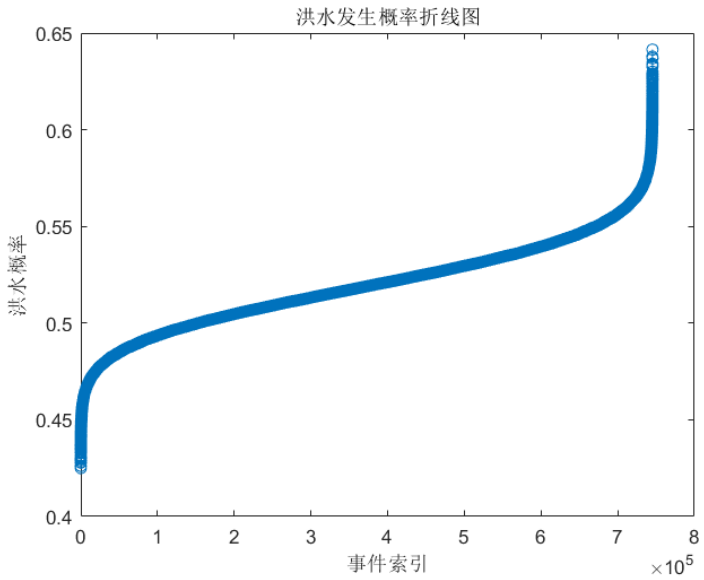

图 7 洪水概率折线图分析

折线图中的数据分布呈现出 S 型曲线，这种形态在接近正态分布的数据中是常见的。通过对折线图的分析，可以更直观地观察到洪水发生概率的数据分布趋势和变化情况。这与直方图的分析结果一致，进一步证明了数据的正态性和集中趋势。

8.3 Kolmogorov-Smirnov 检验

8.3.1 K-S 检验的原理和步骤

Kolmogorov-Smirnov (K-S) 检验是一种非参数统计检验方法，用于比较两个样本分布或一个样本分布与参考分布的差异。K-S 检验的基本思想是通过比较样本的经验分布函数和理论分布函数来判断样本是否符合某个特定的分布。

K-S 检验的步骤：

第一步：假设设定，设定原假设 H0 为样本数据符合某个特定的分布（例如正态分布）。备择假设 H1 为样本数据不符合该特定分布。

第二步：计算经验分布函数，对于样本数据 $X_1,X_2,\ldots,X_n$，计算经验分布函数 $F_n(x)$，即样本数据中小于等于 x 的数据点的比例。

第三步：计算理论分布函数，根据给定的理论分布（本文选用正态分布），计算其分布函数 F(x)。理论分布函数描述了在该分布下数据小于等于 x 的概率。

第四步：计算 K-S 统计量，计算 K-S 统计量 D_n，这是经验分布函数和理论分布函数之间的最大绝对差值。公式如下：

$$D_n = \sup_x \left| F_n(x) - F(x) \right| \tag{8}$$

其中 $\sup_x$ 表示在所有 x 值范围内取最大差值。

第五步：计算 p 值，根据 K-S 统计量 D_n 和样本大小 n，计算 p 值。p 值表示在原假设成立的前提下，观测到当前或更极端统计量的概率。

最后：作出决策，根据显著性水平 α 判断是否拒绝原假设。如果 p 值小于 α，则拒绝原假设，表明样本数据不符合该特定分布。否则，不能拒绝原假设，表明样本数据符合该特定分布。

8.3.2 K-S 检验结果与分析

在对样本数据进行 Kolmogorov-Smirnov (K-S) 检验后，得到以下结果：

表 8 K-S 检验结果

参数名称	计算结果
KS 统计量	0.0362
p 值	0.4234

由于 p 值为 0.4234，大于常用的显著性水平 0.05 或 0.01，因此我们不能拒绝原假设。这意味着，基于 K-S 检验结果，我们没有足够的证据表明样本数据不符

合正态分布。因此，我们可以认为样本数据符合正态分布。

8.4 Q-Q 图

8.4.1 Q-Q 图的基本原理和步骤

Q-Q 图（Quantile-Quantile Plot）是一种用于比较两个分布的工具，特别用于评估一个样本分布是否符合某个特定的理论分布（例如正态分布）。Q-Q 图通过将两个分布的分位数相互比较，以图形化的方式显示两者之间的差异。

Q-Q 图的绘制步骤：

第一步：排序样本数据，将样本数据从小到大排序，得到排序后的样本值 x(1),x(2),…,x(n)。

第二步：计算理论分布的分位数，根据理论分布（本问选用正态分布），计算相应的分位数。对于第 i 个样本值，计算其分位数 q_i，分位数通常是根据理论分布的逆累积分布函数得到的。

第三步：绘制 Q-Q 图，将样本数据的分位数 x(i)作为纵坐标，对应的理论分布的分位数 qi 作为横坐标，绘制散点图。如果样本数据与理论分布相匹配，那么这些点应接近一条 45 度的对角线。

8.4.2 Q-Q 图的绘制结果与分析

对 submit.csv 中的洪水发生概率使用 MATLAB 软件 绘制 Q-Q 图，得到结果如下图。

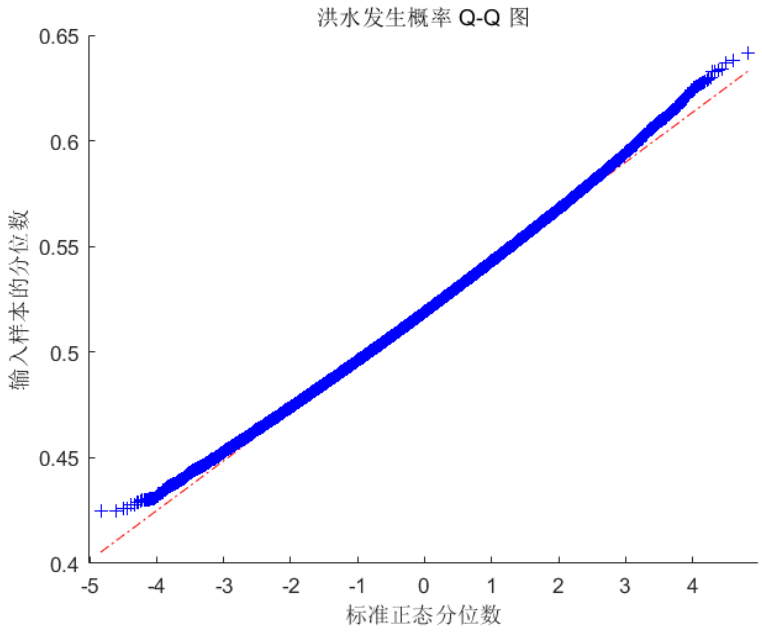

图 8 洪水发生概率 Q-Q 图

从上图可知，样本数据在中心位置上与标准正态分布非常接近，表明其均值和中间范围的分布接近正态分布。在数据分布的极端值区域，样本数据有一些偏离，这表明在尾部，样本数据的分布可能比标准正态分布具有更重或更轻的尾巴。

九、 模型的评价

9.1 模型的优点

优点 1 本文模型结合了皮尔逊相关分析、K-means 聚类、主成分回归、XGBoost 等多种方法，充分利用了数据的特征，保证了预测的准确性和模型的稳健性。

优点 2 在模型建立之前，进行了数据预处理，确保数据无缺失值和异常值，并通过概率分布的可视化分析验证了数据的正态性，这为后续的建模提供了可靠的基础。

优点 3：通过相关性分析和聚类分析，能够明确哪些指标对洪水发生有显著影响，哪些影响较小，并据此提出了合理的预防洪水灾害的建议和措施，具有实际应用价值。

优点 4：通过灵敏度分析，验证了模型对输入数据的微小扰动具有一定的鲁棒性，说明模型在实际应用中能够较好地应对数据的波动和不确定性。

9.2 模型的缺点

缺点 1：模型假设洪水发生概率服从正态分布，但实际情况中，数据可能存在偏态或多峰分布，导致模型在某些情况下预测效果不佳。

缺点 2：在实际应用中，不同数据集可能需要不同的特征选择和参数调优过程，这增加了模型的复杂性和应用难度，需要投入较多的时间和计算资源进行优化。

十、 模型的优化与推广

10.1 模型的优化

优化 1：在现有模型的基础上，可以引入更多与洪水灾害相关的特征变量，如气象预报数据、地质数据、历史洪水记录等，以提高模型的预测精度和适用性。

优化 2：可以尝试使用更加先进的机器学习算法，如深度学习模型（如 LSTM、GRU）或增强学习模型（如 DQN、DDPG），以进一步提升模型的预测能力和鲁棒性。

优化 3：通过集成学习的方法，将多个模型的预测结果进行融合，如采用 Bagging、Boosting 等技术，可以有效降低单一模型的偏差和方差，提高整体预测效果。

10.2 模型的推广

模型不仅可以用于洪水灾害的预测，还可以推广到其他自然灾害的预测与防范中，如地震、台风、滑坡等。通过适当调整和优化特征变量，模型可以适应不同类型灾害的预测需求。

十一、 参考文献

[1]赵琳娜,刘莹,包红军,等.基于集合预报重建法的洪水概率预报研究[C]//中国气象学会.第34届中国气象学会年会 S7 水文气象、地质灾害气象预报理论与应用技术论文集.中国气象科学研究院灾害天气国家重点实验室;成都信息工程大学大气科学学院高原大气与环境四川省重点实验室;四川省气象台;国家气象中心;黑龙江省气象台;,2017:1.

[2]王怡璇,宋松柏.部分概率权重矩在洪水频率分布参数估计中的应用[J].西北农林科技大学学报(自然科学版),2013,41(09):

[3]郑小华,屈振江,栗珂.基于随机过程理论的长江巨洪发生概率预测模式与应用[J].暴雨灾害,2009,28(03):266-270..

[4]施熙灿.水利经济的几种预测方法[J].水利经济,1990,(01):51-59.

[5]展望,李晗冰.农民工身份认同与城市居留意愿——基于主成分分析和随机森林算法的经验证据[J].沈阳工业大学学报(社会科版),2024,17(03):231-241.

十二、 附录

部分核心代码如下：

Matlab 绘图代码

```
 % 获取特征和目标变量
% target = data.FloodProbability; % 替换为实际的目标变量名称
%
% % 获取特征
% features = data(:, 2:end-1); % 从第 2 列到倒数第二列
% featureNames = data.Properties.VariableNames(2:end-1);
%
% % 检查数据缺失值
% if sum(ismissing(data))
%     disp('Data contains missing values, please handle them.');
% else
%     disp('No missing values in the data.');
% end

% 读取数据并保留原始列标题
% data = readtable('train.csv', 'VariableNamingRule', 'preserve');
%
% % 将列名转换为英文
% data.Properties.VariableNames = {'ID', '季风强度', '地形排水', '河流
管理', ...
%     '森林砍伐', '城市化', '气候变化', '大坝质量', '淤积', ...
%     '农业实践', '侵蚀', '无效防灾', ...
%     '排水系统', '海岸脆弱性', '滑坡', '流域', ...
%     '基础设施恶化', '人口得分', '湿地损失', ...
%     '规划不足', '政策因素', '洪水概率'};
%
% % 获取特征名称
% featureNames = data.Properties.VariableNames(2:end-1);
%
% % 绘制直方图
% figure;
% for i = 1:length(featureNames)
%     subplot(5, 4, i); % 5 行 4 列的子图
%     histogram(data.(featureNames{i}), 'Normalization',
'probability');
%
%     title(featureNames{i}, 'Interpreter', 'none');
%     xlabel('值', 'Interpreter', 'none');
%     ylabel('概率', 'Interpreter', 'none');
% end
```

```
% 可视化相关系数
% figure;
% barHandle = bar(correlationMatrix);
% set(gca, 'XTickLabel', chineseLabels, 'XTickLabelRotation', 45);
% ylabel('相关系数', 'FontName', 'SimHei');
% title('特征与洪水概率的皮尔逊相关系数', 'FontName', 'SimHei');
% grid on;
%
% % 在每个柱状图上添加数值标签
% xtips = barHandle.XEndPoints;
% ytips = barHandle.YEndPoints;
% labels = string(barHandle.YData);
% text(xtips, ytips, labels, 'HorizontalAlignment', 'center',
'VerticalAlignment', 'bottom', 'FontName', 'SimHei');
%
% % 设置字体和图形外观
% set(gca, 'FontSize', 12, 'FontName', 'SimHei'); % 使用黑体字体
% xlabel('特征', 'FontName', 'SimHei');
% ylabel('相关系数', 'FontName', 'SimHei');
% title('特征与洪水概率的皮尔逊相关系数', 'FontName', 'SimHei');
%
% % 调整图形大小和布局
% set(gcf, 'Position', [100, 100, 1500, 800]); % 设置图形大小
% 提取特征和目标变量
% features = data{:, 2:end-1}; % 所有特征（不包含 ID 和洪水概率）
% target = data.FloodProbability; % 目标变量
%
% % 标准化处理
% features = zscore(features);
%
% % 主成分分析
% [coeff, score, ~, ~, explained] = pca(features);
%
% % 选择前几个解释方差较高的主成分
% numComponents = 5; % 根据需求选择前几个主成分
% selectedComponents = score(:, 1:numComponents);
%
% % 使用主成分进行线性回归
% model = fitlm(selectedComponents, target);
%
% % 显示回归模型结果
% disp(model);
%
% % 获取主成分的权重
```

```
% pcWeights = model.Coefficients.Estimate(2:end);
%
% % 创建一个表格保存主成分名称和权重
% pcNames = {'PC1', 'PC2', 'PC3', 'PC4', 'PC5'};
% pcWeightsTable = table(pcNames', pcWeights, 'VariableNames',
{'PrincipalComponent', 'Weight'});
%
% % 保存到 Excel 文件
% writetable(pcWeightsTable, 'principal_component_weights.xlsx',
'WriteVariableNames', true);
%
% % 输出主成分的权重
% disp('主成分的权重：');
% disp(pcWeightsTable);
%
% % 获取主成分的计算公式和权重
% principalComponentFormulas = coeff(:, 1:numComponents);
%
```