**一、课题综述**

* 1. **课题说明**

实验分工如下，报告内容共同完成：

* 1. **课题目标**

本实验针对中国台湾省各地区的空气质量数据集建立多种回归模型，以预测空气质量指数（AQI），并对比分析模型预测效果差异，为空气污染防治措施提供数据支持和决策依据。数据集包含多维度的空气污染物浓度、气象条件和地理信息等特征。

数据集链接：https://www.kaggle.com/datasets/taweilo/taiwan-air-quality-data-20162024

* 1. **课题数据集**

数据集来源于Kaggle，包含2016年至2024年期间中国台湾省各地区空气质量监测数据。数据集共1048575个样本，包含多维度信息：

1. 空气污染物浓度：如PM2.5、PM10、SO2、CO、O3、NO2等主要污染物的实时监测数据，反映了空气中不同污染物的浓度水平
2. 气象条件：风速、风向等相关气象因素
3. 时间信息：数据按小时收集，涵盖了2016年至2024年间的时间跨度，包含年份、月份、日期和小时等信息，便于分析空气质量的时序变化。
4. 地理信息：数据集提供了监测站点的地理位置，包括站点所属的县市和经纬度坐标，便于按区域对空气质量进行对比分析。

**二、实验报告设计**

* 1. **数据准备**

**2.1.1 数据集获取**

数据集于2024年9月在[Kaggle](https://www.kaggle.com/datasets/taweilo/taiwan-air-quality-data-20162024)发布，保证了课题的创新性

**2.1.2 数据加载**

使用pandas库将数据集以CSV形式加载到Python3环境

* 1. **数据预处理**

**2.2.1 特征选择（初步）**

去除数据集中除aqi以外的标签列status，去除缺失值超过80%的unit列，确保特征有足够有效数据支撑，剩余1列lable和19列特征

**2.2.2 数据类型转换**

将所有数值型特征统一转化为float，分类特征pollutant保留object类型

**2.2.3 异常值处理**

监测数据中存在一些极端值或异常值（如污染物浓度过高或过低），误导模型训练和预测。使用IQR方法剔除异常样本，提升数据可靠性。

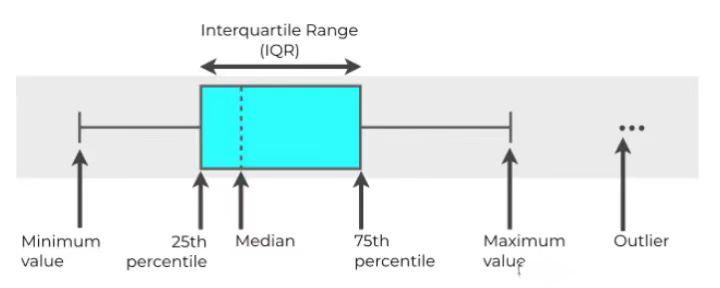


图 2 IQR方法

**2.2.4 缺失值处理**

Pollutant类别特征缺失使用NAN填充，其余数值特征如有缺失，删除缺失数据所在行

**2.2.5 特征编码**

对分类特征pollutant创建one-hot虚拟变量，将其数值化

**2.2.6 特征标准化**

为避免特征之间的数值范围差异过大导致模型在训练时产生偏差，对特征（除aqi）进行Z-SCORE标准化，转换成具有零均值和单位方差的标准正态分布形式，使模型更好地收敛。

**2.2.7 特征选择（PCA）**

随机森林回归模型中，我们尝试采用主成分分析进行特征选择，保留了能解释数据90%方差的14个主成分进行模型训练和评估。结果显示，模型的R²评分为0.894，MSE为55.73，平均MAE为5.43，相较不使用PCA预测效果反而下降，原因可能有以下2点：

PCA 假定特征线性相关，通过线性组合来表示原始特征。而数据集特征之间的特征可能是非线性的，且对回归任务较为重要

PCA 的目的是降维，舍弃原始数据的一部分信息。它尽量保留了最大的方差，但可能并不完全对应于回归任务中最重要的特征，可能影响回归效果。

**2.2.8 数据集划分**

训练集和测试集按3：2随机划分，分别用于模型训练和验证泛化能力。

通过以上数据预处理，我们将原始数据转化为一个清晰、规范且易于模型处理的数据集。在整个数据预处理过程中，清理缺失值和异常值确保了数据的质量，标准化保证了数据的一致性，这些步骤为后续回归模型的训练和预测奠定了坚实基础，使结果具有更高可信度和可解释性。

* 1. **模型搭建**

**2.3.1 模型评估指标**

* 均方误差 （MSE）：
* 决定系数（R²）：
* 平均绝对误差 (MAE)：

**2.3.2 线性回归模型**

线性回归模型尝试寻找一个线性方程，来表示自变量（输入特征）与因变量（AQI）之间的关系。具体形式为：

N个样本的矩阵形式：

* 拟合模型：通过调用fit函数，模型根据训练数据进行参数调整，找到最佳的系数组合。

**2.3.3 套索回归和岭回归模型**

在多元线性回归中，当自变量存在多重共线性时，传统最小二乘估计（OLS）方法可能会导致模型不稳定，且预测性能差。因此采用正则化回归方法，Lasso回归和Ridge回归是两种常用的正则化技术。

* 回归模型的预测公式：
* Lasso回归通过添加L1正则化项实现特征选择，提高模型可解释性:
* Ridge回归通过添加L2正则化项减少模型的复杂性，防止过拟合:

**2.3.4 神经网络回归模型**

前馈神经网络（FNN）是最基本的人工神经网络模型，其特点是信息流在网络中单向从前向后传递，没有反馈连接，整个过程中不存在环路。实验过程中发现简单的网络回归效果已经较好，故不再选择更复杂的网络结构。

* 基本结构：

输入层：全连接层，含64个节点，激活函数为ReLU，输入数据的特征数量为21。

隐藏层：全连接层，含64个节点，激活函数为ReLU。

输出层：全连接层，含1个节点，用于回归任务，预测一个连续值输出。

* 优化器：RMSprop

学习率初始设为0.001。RMSProp 优化器是基于梯度下降的优化算法，主要思想是根据梯度的平方的衰减平均值来调整参数。与传统的梯度下降不同，RMSProp 在每一次迭代中会对梯度的平方进行衰减，从而避免梯度爆炸和消失问题，同时调整学习率来控制优化速度。

* 损失函数：均方误差（MSE）

模型还记录平均绝对误差（MAE）和均方误差（MSE）以观察性能

**2.3.4 随机森林回归模型**

随机森林回归模型是由多个决策树组成的集合模型，其基本结构的搭建主要包含以下几个步骤：

* 数据抽样：

使用Bagging方法数据采样

* 回归树的构建与训练：

对每个训练集训练一颗回归树。在构建树中每个节点分裂时，只从一部分特征中选取最佳特征进行分裂，从而增加模型的差异性，减少了各树之间的关联。

* 集成预测：

对每一个新的输入样本，随机森林的每颗回归树都会给出一个预测值Ti(x)，最终的预测结果是所有树预测值的平均值

* 1. **模型训练测试**

**2.4.1 朴素线性回归、套索回归和岭回归**

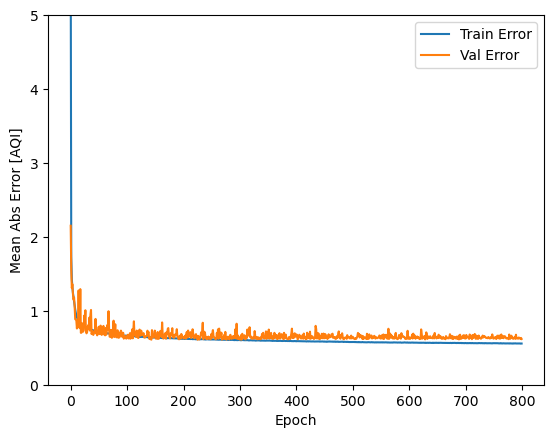
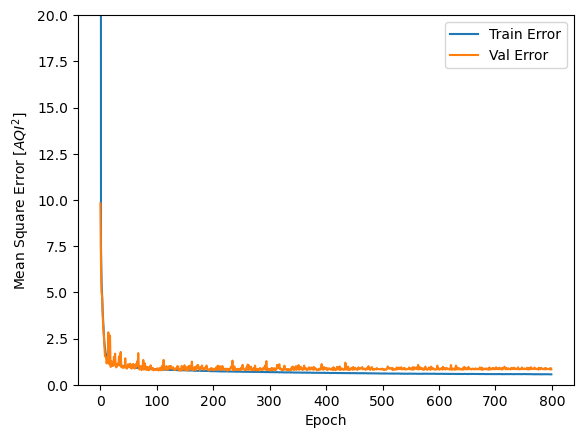
训练参数及训练集上最终各项指标：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型\参数 | 正则化参数α | MSE | MAE |  |
| 朴素线性回归 | \ |  |  |  |
| Lasso回归 | 0.1 | 33.4492 | 4.013 | 0.9367 |
| 岭回归 | 0.1 | 32.7605 | 4.019 | 0.9380 |

**2.4.2 神经网络回归**

经过多次调整，训练epoch最终设为100，验证集的划分比例为0.4

* 训练过程中训练集和验证集MAE（AQI)、MSE（AQI)随Epoch的变化曲线：

* 最后几轮训练过程中的各项指标：

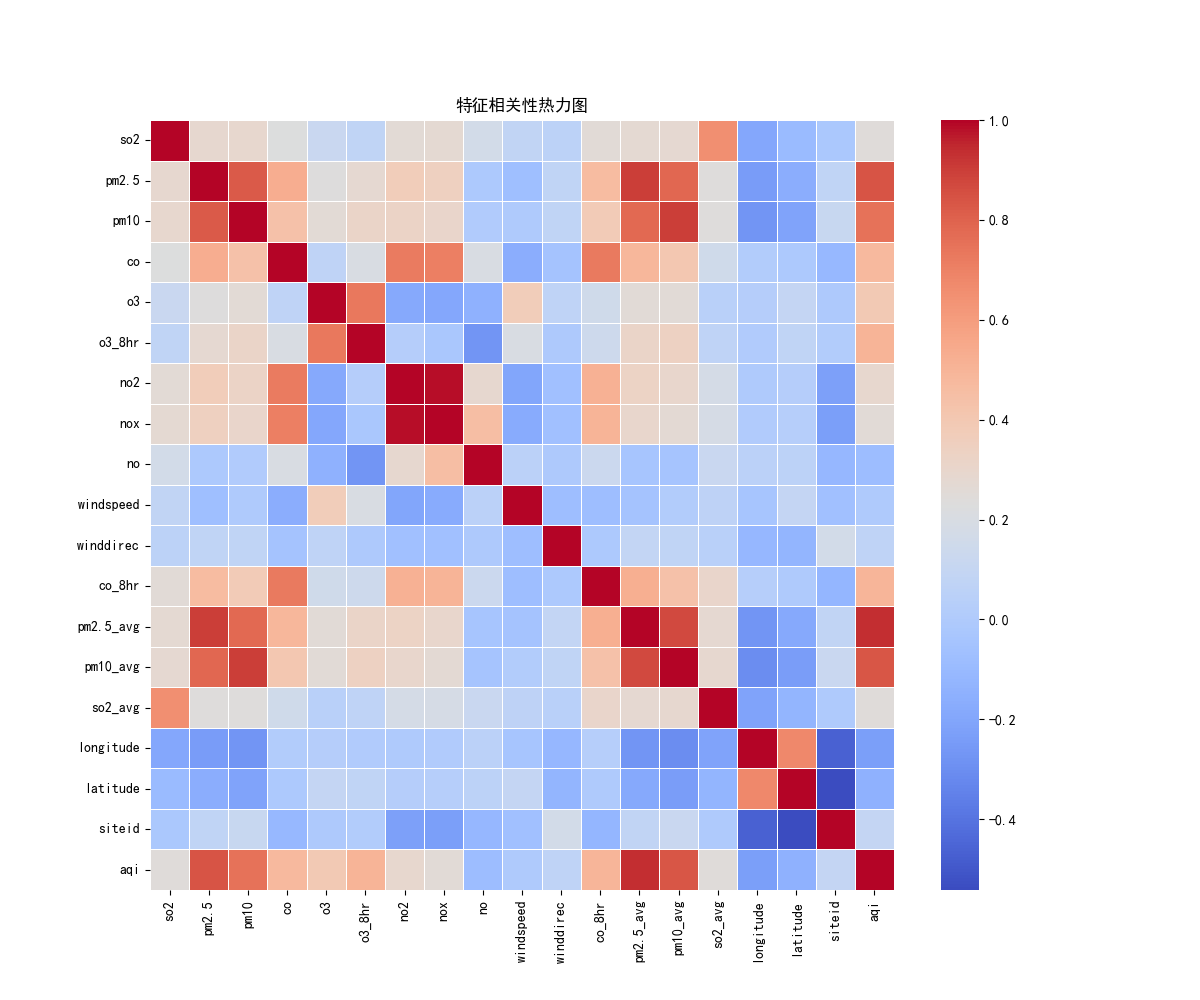


**2.4.3 随机森林回归**

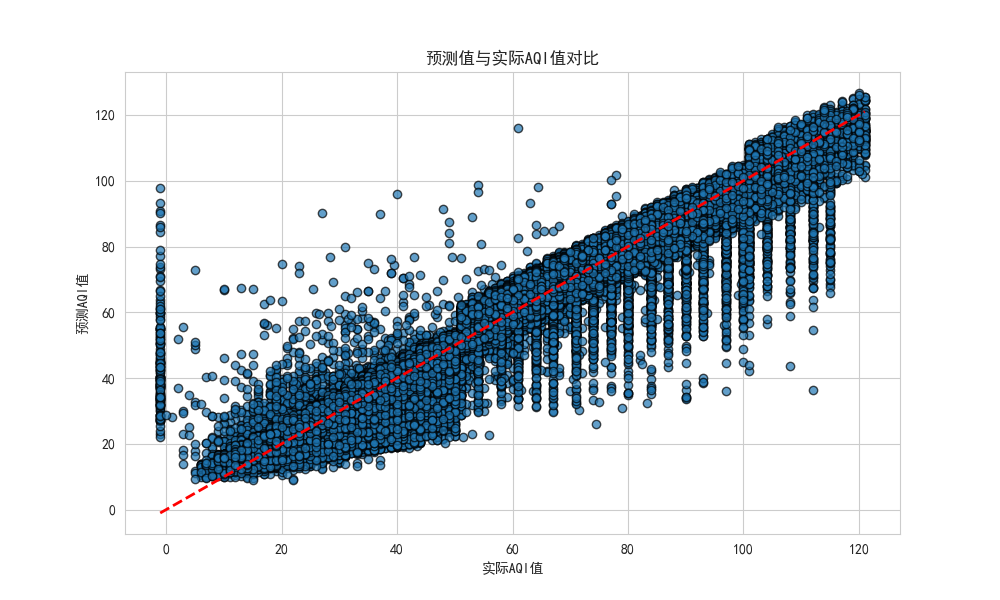
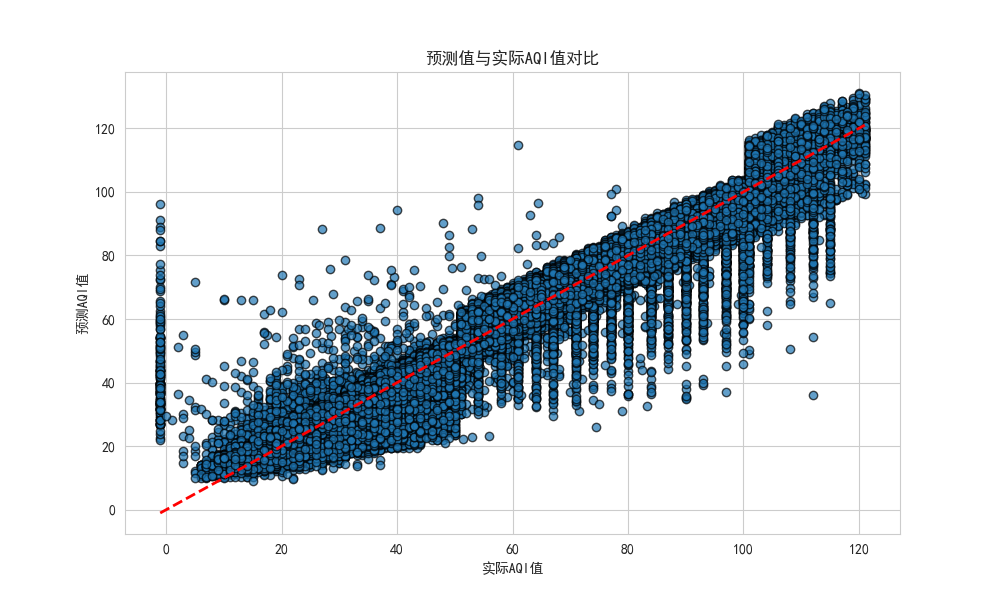
* 1. **结果可视化**

各模型AQI预测值与真实值可视化如下：

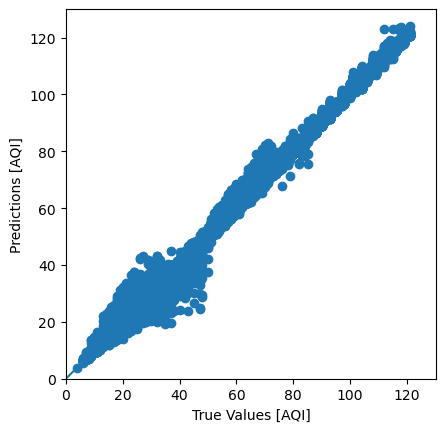
* 相关特征热力图：



* 朴素线性回归：
* Lasso回归（上）和岭回归（下）

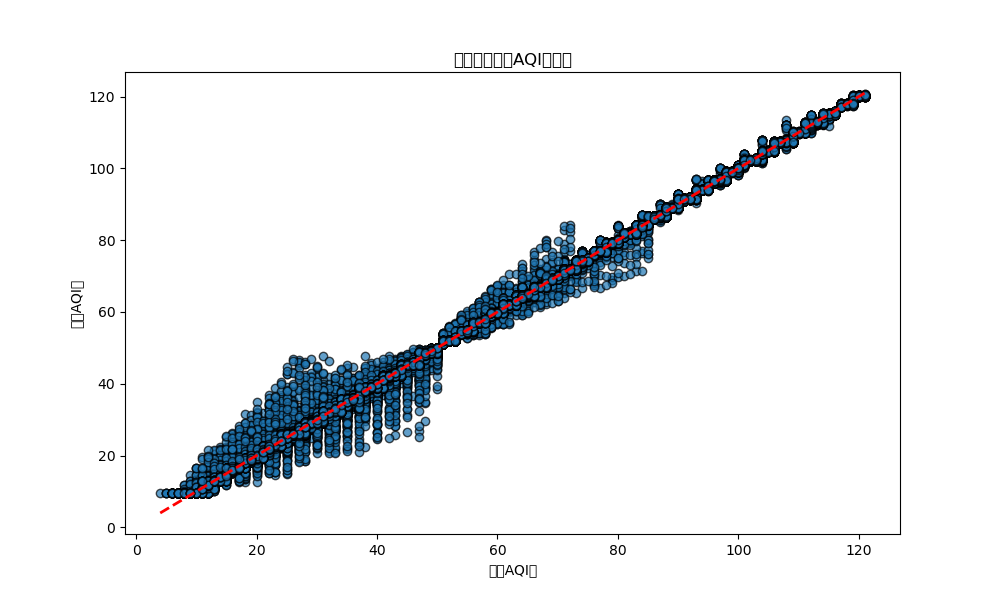
 

* 神经网络：



另外，神经网络模型测试集上的损失函数值为0.9751，损失值较低，说明模型泛化能力较好。

* 随机森林：



以上使用散点图直观展示各模型预测值与真实值的关系，结果越接近直线

，预测值与真实值之间的一致性越好，即预测准确度高

* 1. **分析和优化**（要包含对两类模型的结果的比较讨论）

**2.6.1 模型结果对比分析**

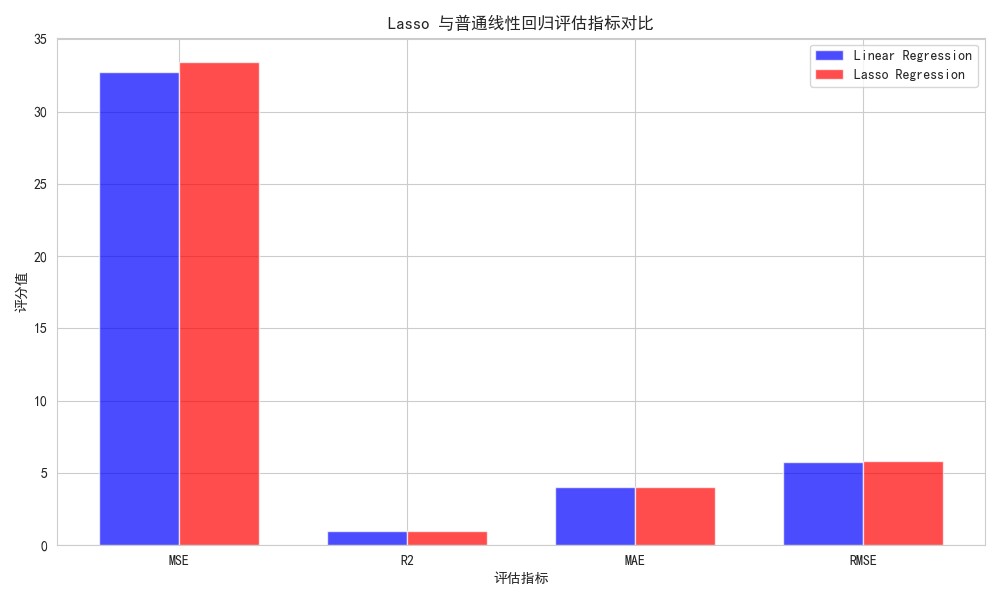
测试集上最终各项评价指标：

|  |  |  |  |
| --- | --- | --- | --- |
| 模型\test指标 | MSE | MAE |  |
| 朴素线性回归 |  |  |  |
| Lasso回归 | 33.4493 | 4.013 | 0.9367 |
| 岭回归 | 32.7605 | 4.019 | 0.9380 |
| 随机森林 | 0.6317 | 0.4556 | 0.9988 |
| 神经网络 | 0.8711 | 0.6401 | 0.9983 |

* 测试集上的平均绝对误差（MAE）是模型在测试集上的预测值与真实值之间的平均绝对差异，MAE越小意味着模型的预测误差相对较小
* 测试集上的均方误差（MSE）是模型在测试集上的预测值与真实值之间的差异的平方平均值，由于MSE对较大的误差敏感，因此更能反映模型在预测较大误差时的表现。
* 表明模型中的特征能够解释AQI 指数多少的变化，越接近1表示特征与AQI指数之间越有很强的相关性，还意味着预测误差较小，有效性好。

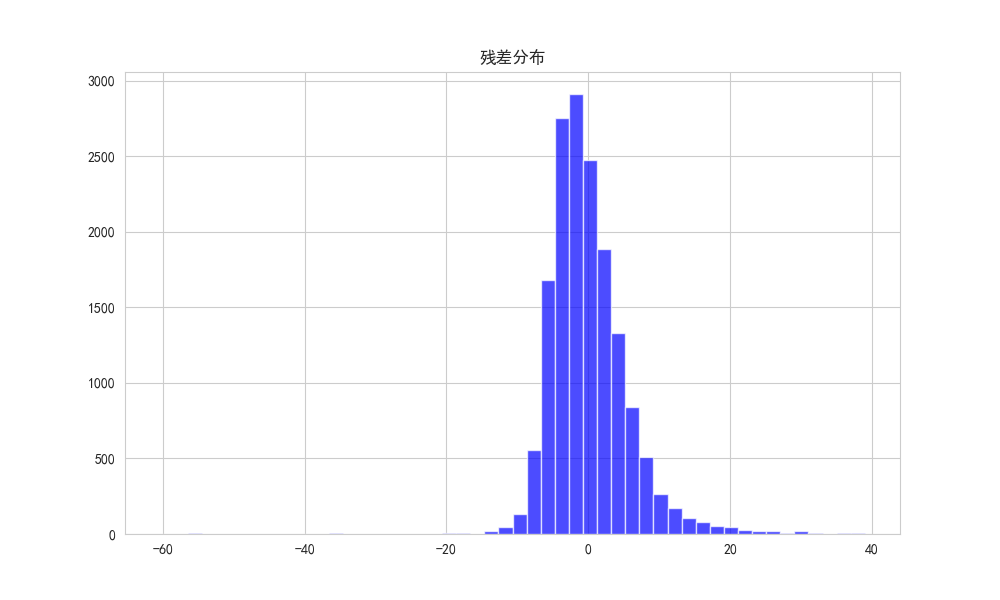
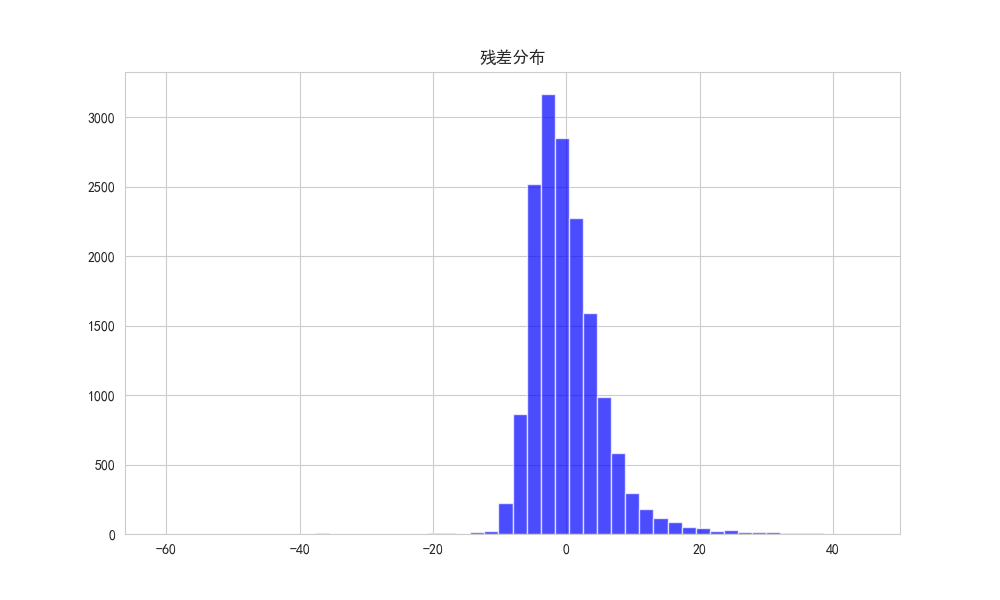
可知随机森林和神经网络的预测效果较好，各模型的结果分析如下：

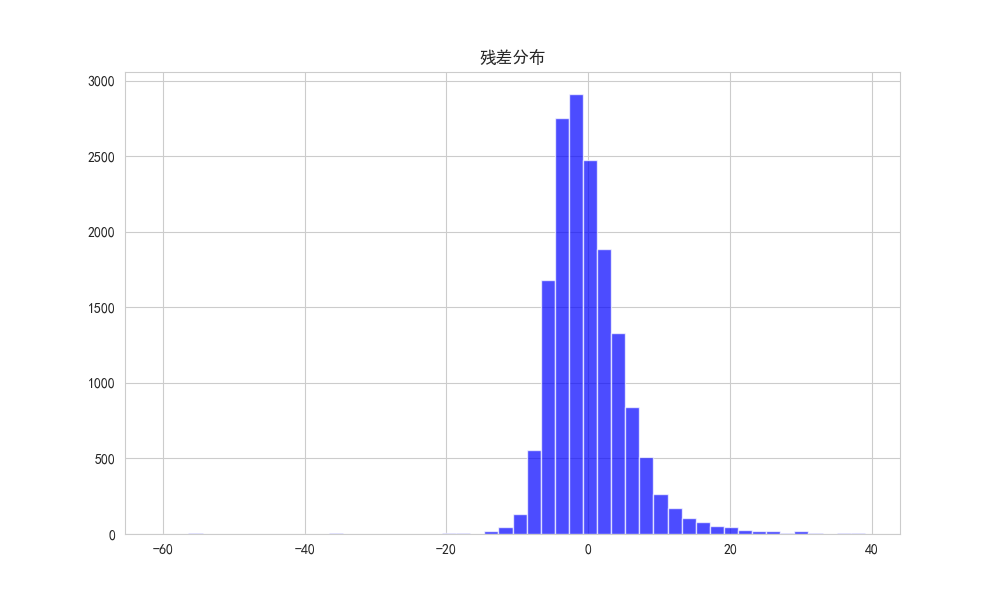
**2.6.2 线性回归、套索回归及岭回归结果比较**



图中可以看出Lasso回归在MSE、MAE和RMSE上的表现均略好于线性回归，但在R2上略逊一筹。

* 残差分布（线性回归，lasso回归，岭回归）：



三种模型的残差表现基本一致，没有显著的优劣之分。

对套索回归和岭回归分析：

* 指标解析

MSE:岭回归的MSE（32.7605）小于套索回归（33.4493），表明岭回归在预测AQI方面稍有优势

R²评分:两个模型的R²值都接近0.94，说明都能很好地解释数据变异性。

MAE:套索回归的MAE（4.0129）比岭回归（4.0195）低，表明套索回归在此方面表现稍好

* 模型对比

模型选择:若更注重整体预测精度，可能更倾向于选择岭回归；若更关注单个预测的准确性，则可以考虑套索回归。

正则化效果: 套索回归会对某些特征进行压缩（设为零），从而实现特征选择，适合高维度数据；而岭回归则对所有特征施加均匀惩罚，更适合多重共线性问题。依据数据的特性和需求，可以选择不同的正则化方法。

适用场景: 如果数据中存在许多不重要的特征，套索回归可能更合适；如果数据特征之间存在共线性，岭回归可能会更稳定。

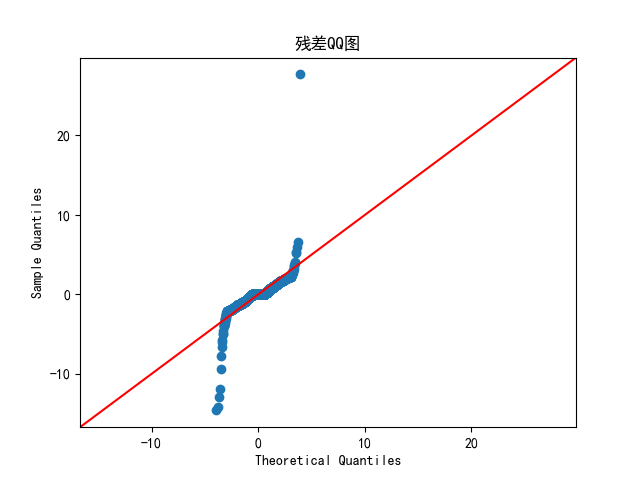
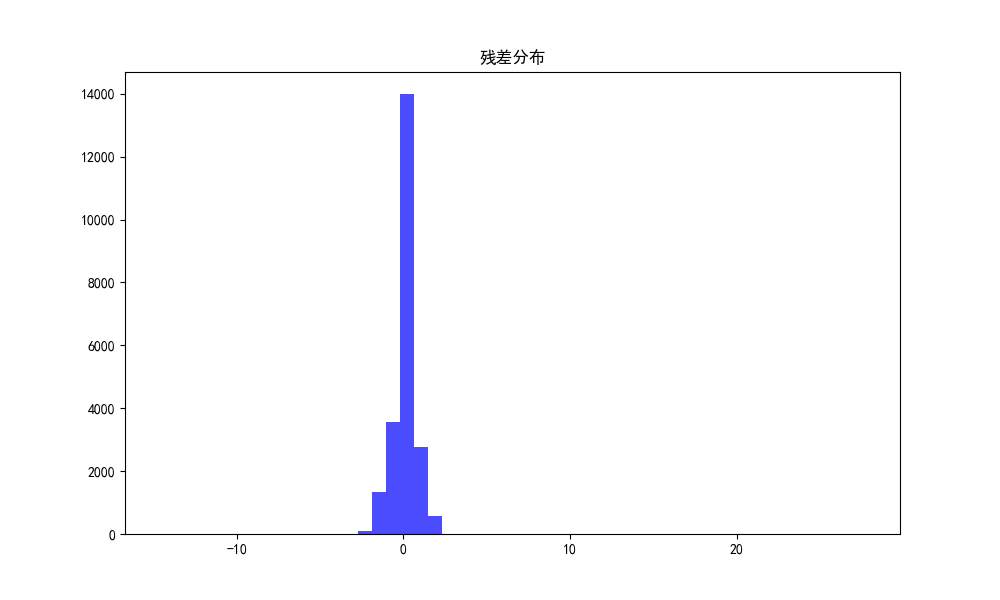
* 模型优化

交叉验证: 不同的正则化强度会影响模型的复杂性和预测能力， 可以使用K折交叉验证选择最佳的正则化参数。

网格搜索和随机搜索: 可使用网格搜索或随机搜索探索参数空间，从而找到最佳的正则化参数组合。

**2.6.3 随机森林结果分析与优化过程**

* 残差分布图、残差QQ图：



* 结果分析与模型优化

1. **总结**

要求说明：在提交实验报告时请**将代码以附件形式随报告一起提交**

**实验报告的格式无限制，页数限制在10页内**