浙江農林大學
ZHEJIANG A&F UNIVERSITY

机器学习

学生姓名：黄豪

学　　号：2024611031011

专业班级：数计研 241 班

所在学院：数学与计算机科学学院

浙江農林大學

硕士生课程论文（设计）诚信承诺书

我谨在此承诺：本人所写的课程论文（设计）《回归作业报告》均系本人独立完成，没有抄袭行为，凡涉及其他作者的观点和材料，均作了引用注释，如出现抄袭及侵犯他人知识产权的情况，后果由本人承担。

承诺人（签名）：黄豪

2024 年 12 月 21 日

回归作业报告

数学与计算机科学学院　数计研 241　黄豪　　指导教师：夏凯

摘要：本文以混凝土抗压强度数据集为研究对象，探讨了机器学习算法在回归任务中的性能表现。通过分析 8 个输入特征与目标变量抗压强度之间的关系，构建并评估了多种回归模型，包括规则树-线性回归、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机和极限梯度提升。实验结果表明，轻量级梯度提升机和极限梯度提升模型在预测精度和泛化能力方面表现最佳，分别达到了最低的均方根损失（RMSE=0.2952 和 0.3006）和较高的决定系数（R^2=0.9225 和 0.9196），此外变量相关性分析揭示了水泥含量和龄期对抗压强度的显著正相关性，提供了优化混凝土配方的重要参考依据。本文的研究不仅验证了多种机器学习模型在回归任务中的适用性，还为实际工程中混凝土性能的精准预测和材料优化提供了科学指导。

更多详细工作内容尽在：https://github.com/H-0526/Machine_Learning/tree/master

关键词：机器学习；集成学习；回归；随机森林

Regression Assignment Report

Abstract: This study focuses on the concrete compressive strength dataset to explore the performance of machine learning algorithms in regression tasks. By analyzing the relationship between 8 input features and the target variable (compressive strength), various regression models were constructed and evaluated, including Cubist regression, Deep Random Forest, Extremely Randomized Trees (Extra Trees), Gradient Boosted Decision Trees (GBDT), Light Gradient Boosting Machine (LightGBM), Random Forest, Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). The experimental results indicate that LightGBM and XGBoost demonstrated the best performance in terms of prediction accuracy and generalization ability, achieving the lowest root mean squared errors (RMSE=0.2952 and 0.3006) and high R-squared values (R^2=0.9225 and 0.9196), respectively. Additionally, the variable correlation analysis revealed significant positive correlations between cement content, age, and compressive strength, providing critical insights for optimizing concrete mixtures. This study not only validates the applicability of various machine learning models in regression tasks but also offers scientific guidance for accurate prediction of concrete performance and material optimization in practical engineering applications.

For more details, visit: https://github.com/H-0526/Machine_Learning/tree/master

Key words: Machine Learning, Ensemble Learning, Regression, Random Forest

目　录

1 绪论

1.1 数据集介绍

混凝土抗压强度数据集 (Concrete Compressive Strength Data Set) 是一个常见的回归分析任务数据集，广泛用于机器学习和数据挖掘领域的研究与教学。该数据集来源于 UCI 机器学习数据集仓库，包含 1030 个样本，记录了混凝土成分及其抗压强度之间的关系。数据集提供了 8 个输入特征，包括水泥含量（kg/m^3）、炉渣含量（kg/m^3）、飞灰含量（kg/m^3）、水含量（kg/m^3）、减水剂含量（kg/m^3）、粗骨料含量（kg/m^3）、细骨料含量（kg/m^3）和混凝土龄期（天），以及一个目标变量抗压强度（MPa），代表混凝土在不同配比和养护龄期下的强度。

该数据集适合探索混凝土抗压强度与其成分之间的关系，涉及特征选择、数据预处理、模型优化等步骤。经分析，回归任务将以预测抗压强度为目标，此外利用相关性分析与特征重要性评估，可以深入揭示不同成分对抗压强度的影响，为混凝土配方设计提供科学依据。在模型评估方面，可通过 MSE、RMSE、Mean CV score、MAE、R^2等指标对模型性能进行综合比较，结合可视化手段，该数据集还支持更直观地理解数据分布及模型预测效果。这些分析不仅为研究混凝土力学性能提供了理论支持，也对建筑工程中的材料设计与优化具有实际指导意义。

1.2 数据集理解

混凝土抗压强度数据集 (Concrete Compressive Strength Data Set) 是一个典型的回归问题数据集，适用于探索混凝土成分与其抗压强度之间的定量关系，包含 8 个特征变量和 1 个目标变量抗压强度(MPa)。该数据集适合进行回归分析、特征工程、相关性分析和模型性能评估等任务，常用于验证线性回归、随机森林回归和神经网络等模型的效果。通过对成分变量的重要性分析，可以揭示各成分对混凝土性能的影响，为材料设计与优化提供数据支持。尽管数据来源于实验环境，可能面临非线性关系复杂、变量交互作用强和分布不均等挑战，但其在材料科学与机器学习领域均具有重要研究价值，为实际工程中的混凝土配方优化和性能预测提供了科学依据。

1.3 研究目的与内容

1.3.1 研究目的

本研究旨在探索机器学习算法在混凝土抗压强度预测任务中的应用，以提高对混凝土性能的精准预测能力。通过系统评估多种回归模型的表现，包括规则树-线性回归、集成学习模型（如随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、极限梯度提升）以及支持向量机等，研究不同算法在处理非线性特征、变量相关性和复杂数据分布时的优劣。通过分析数据分布特性、变量相关性和异常值，优化特征选择和数据处理流程，确保模型构建的科学性和有效性。本研究不仅希望验证集成学习算法在回归任务中的优势，还旨在为混凝土配方优化和工程应用中的性能预测提供数据支持，为进一步的模型优化和算法创新奠定基础。

1.3.2 研究内容

第一部分：绪论。绪论部分概述了数据集的来源并给与介绍，明确了混凝土抗压强度回归任务，同时本部分还阐述了研究的目的与内容，给出了研究的技术路线

第二部分：相关回归方法介绍。本部分系统地介绍了研究中采用的多种回归方法，包括规则树-线性回归、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机和极限梯度提升等经典模型，此外还探讨了不同模型在回归任务中的理论优势及其适用场景。

第三部分：基于机器学习算法的实验过程。本部分是研究的核心内容，详细描述了实验的设计与实施过程，包括数据预处理、特征选择、模型训练与优化、以及评价指标的选定。实验对比了多种机器学习模型在混凝土抗压强度数据集上的性能表现，并结合回归实验结果，对模型适用性进行了深入分析。

第四部分：总结与展望。研究最后对实验结果进行了总结，指出了不同模型的优劣势及其对混凝土抗压强度回归任务的启示。同时对本研究的不足之处进行反思，并展望了未来在回归方法改进、数据增强及应用领域扩展方面的潜在研究方向。

1.3.3 技术路线

本文的技术路线如图 1-1 所示。

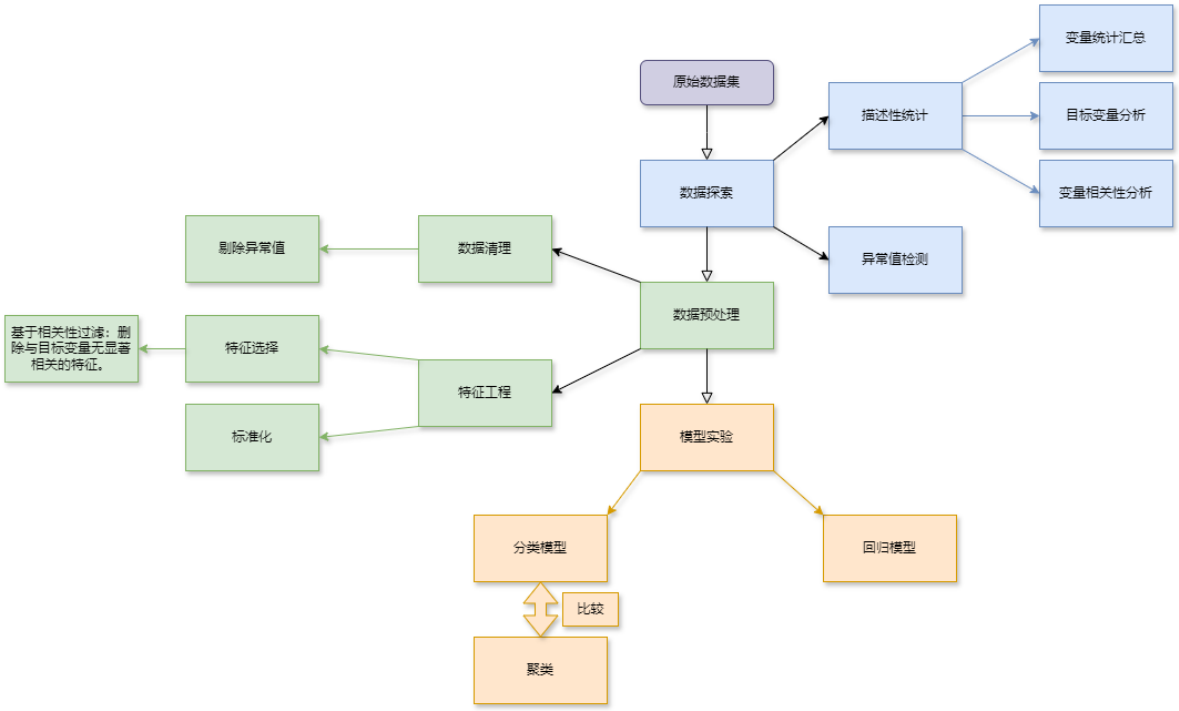

图 1-1 技术路线图

2 相关回归方法介绍

2.1 规则树-线性回归

Cubist 模型是一种结合回归树和线性回归的预测模型，专为处理回归任务而设计，其核心思想是通过构建一组规则树并结合局部线性模型实现精准预测。与传统回归树相比，Cubist 不仅通过分裂特征空间生成树结构，还在每个叶节点内拟合线性回归模型，以提高对连续型目标变量的拟合能力，Cubist 模型能够捕捉全局非线性特性和局部线性关系，是一种高效且灵活的回归方法。

模型的预测可以表示为以下形式：

$$\hat{y} = \sum_{i=1}^{T} w_i \cdot g_i(x) \tag{2.1}$$

其中，T 表示规则树的数量，$g_i(x)$是第 i 棵规则树的预测值，w_i 是对应的权重。每棵规则树通过分裂特征空间生成一组叶节点，在每个叶节点内，使用局部的线性回归模型来预测目标变量，从而形成一种混合模型。Cubist 在预测过程中首先根据输入样本 x 寻找到对应的规则路径，依次计算每棵规则树的线性回归输出 $g_i(x)$，然后通过加权求和得到最终预测值。权重 w_i 通常根据每棵规则树的适应性进行调整，以提升整体的预测性能。

与传统回归树或随机森林不同，Cubist 不仅能对非线性数据进行建模，还能通过局部线性模型提高对连续目标变量的精细刻画能力。该模型在数据包含复杂特征交互关系时表现尤为优越，是一种兼具灵活性和准确性的回归工具，广泛应用于环境建模、经济预测和工程问题中。

2.2 深度随机森林

利用多层神经网络（通常是全连接层）对原始特征进行嵌入变换，生成高维、抽象的特征表征。

特征嵌入的函数形式如下：

$$h(x) = f(Wx + b) \tag{2.2}$$

其中 W 和 b 分别为神经网络的权重和偏置，f(·)是激活函数。

之后是随机森林预测，公式如下：

$$\hat{y_i} = \frac{1}{T}\sum_{t=1}^{T} f_t\big(h(x_i)\big) \tag{2.3}$$

其中 T 是森林中树的数量，$f_t(\cdot)$表示第 t 棵树的预测结果。

深度随机森林通过结合神经网络的特征提取能力和随机森林的高效回归能力，在回归任务中表现优异。同时通过联合优化策略，深度随机森林能够在不显著增加模型复杂度的情况下提升预测性能，兼顾深度学习的表达能力和传统方法的高效性，是回归任务中的一个重要方向。

2.3 极端随机树

Extra-Trees 属于基于树的集成学习方法，与随机森林类似，但在分裂点选择和样本使用上引入了更高的随机性，其目标是构建一组多样化的决策树，设输入特征为 $x=[x_1,x_2,\ldots,x_d]^T$，目标变量为 y，Extra-Trees 的目标是构建 T 棵极端随机树，通过统计取平均值得到预测结果，具体公式如下。

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \tag{2.4}$$

其中 $f_t(x)$是第 t 棵树对样本 x 的预测结果。

2.4 梯度提升决策树

梯度提升决策树（GBDT）是提升方法（Boosting）的一种实现，通过迭代训练一组决策树，将每棵树的错误修正为下一棵树的目标，逐步逼近目标变量 y 的真实值，从而实现高精度的预测。

GBDT 将模型表示为一组累加的决策树：

$$\mathrm{F}(x) = \sum_{m=1}^{M} f_m(x) \tag{2.5}$$

其中 M 是决策树的数量，$f_m(x)$是第 m 棵树的预测值，F(x)是模型的最终预测值，通过初始化、拟合伪残差、构建弱学习器、更新模型、迭代重复，最终计算得到模型输出：

$$\hat{y} = \mathrm{F}(x) = \sum_{m=1}^{M} f_m(x) \tag{2.6}$$

2.5 轻量级梯度提升机

轻量级梯度提升机（LightGBM）是一种基于梯度提升框架的高效算法，专注于在大规模数据集和高维特征场景下提升训练速度与内存效率，同时保留梯度提升决策树（GBDT）的强大性能。LightGBM 在回归任务中的基本原理与分类任务类似，但其优化目标和模型表示方式针对回归问题进行了调整。

LightGBM 与 GBDT 的基本原理相同，旨在通过迭代优化逐步逼近目标函数。在回归任务中，LightGBM 通过将连续特征离散化为 G 个区间，生成直方图表示特征分布，然后在每个区间计算梯度和二阶梯度的累积值，从而快速找到最优分裂点，与传统 GBDT 的主要区别在于其对数据处理和树生长的优化。首先 LightGBM 引入了基于直方图的决策树构建方法，通过将连续特征离散化为有限个区间，大幅降低了分裂点搜索的复杂度，先对每个特征 x_j，将取值划分为 G 个区间。再在直方图上计算每个区间的梯度和损失，选择最优分裂点，其优化公式为：

$$Split\ Gain = \frac{(G_{left})^2}{H_{left} + \lambda} + \frac{(G_{right})^2}{H_{right} + \lambda} - \frac{(G_{total})^2}{H_{total} + \lambda} \tag{2.7}$$

其中 G 和 H 分别表示梯度和二阶梯度累积，λ 为正则化项。传统 GBDT 使用按层生长的树结构，而 LightGBM 采用基于叶子的生长策略，优先分裂能带来最大增益的叶子节点，目标是最大化信息增益。

$$Leaf\ Gain = \frac{(G_{leaf})^2}{H_{leaf} + \lambda} \tag{2.8}$$

这一策略能更有效地适应数据分布，提高回归任务的性能。

2.6 随机森林

随机森林模型是一种集成学习算法，综合多个弱学习器的输出，统筹做出预测，一般以决策树为弱学习器，决策树的构造涉及特征选取、构造决策树、剪枝三步。随机森林在每个决策树的节点分裂时随机选择部分特征进行评估，常用的特征选取方法包括 ID3、C4.5、CART。

信息增益（ID3）衡量特征对样本信息熵的降低程度，信息增益更适合处理离散特征，但对多值特征有偏好，公式为：

$$g(D|A) = H(D) - H(D|A) \tag{2.9}$$

信息增益率（C4.5）修正信息增益对多值特征的偏好，引入了特征值熵 H(A)，其

公式为：

$$g_r(D|A) = \frac{g(D|A)}{H(A)} \tag{2.10}$$

基尼指数（CART）：衡量样本集合的纯度，分裂节点时选择基尼指数最小的特征，其公式为：

$$Gimi(D) = 1 - \sum_{k=1}^{K} P_k^2 \tag{2.11}$$

随机森林是集成学习 Bagging 的变体，训练过程中通过随机选择样本和属性，因此可以发现随机森林的弱学习器差异性及来自样本和属性，可提升泛化能力。随机森林使用自助采样法，从包含 m 个样本的数据集 D 中进行有放回的抽样 m 次，形成新的数据集 D^*。当样本数量趋于无穷，可计算出样本未被采集的概率：

$$lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m = \frac{1}{\mathrm{e}} \approx 0.368 \tag{2.12}$$

可看出元数据及中将有 36.8%的样本未被采用，从未被用来训练，这将给基学习器带来较大差异，随机森林模型的目标是通过构建多个回归树并结合其预测结果，再对连续型目标变量进行准确预测提高准确性。

2.7 支持向量机

支持向量回归（SVR）是支持向量机（SVM）在回归任务中的具体应用，其目标是找到一个函数，使其偏离目标值的误差不超过给定的阈值，同时尽可能保持模型的平滑性。SVR 通过引入 ϵ-不敏感损失函数，在不计入小误差的情况下拟合数据，避免对噪声的过度敏感。

SVR 的拟合的目标函数为：

$$f(x) = w^T x + b \tag{2.13}$$

其中 w 是法向量，表示函数的方向；b 是偏置，用来控制函数与原点的距离，而为了处理非线性数据或允许偏离目标值 y_i 的误差不超过 ϵ，引入松弛变量 ξ，优化目标变为：

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{2.14}$$

其中 C 是正则化参数，用于控制模型对误差和复杂度的权衡，而当数据无法线性拟合时，SVM 使用核函数将数据映射到高维特征空间，使其在高维空间中线性拟合，

常用的高斯核函数为：

$$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \tag{2.15}$$

支持向量回归（SVR）通过引入 ϵ-不敏感损失函数，实现对连续目标变量的高效回归预测，其核心特点是在控制模型复杂度的同时尽量逼近目标值，忽略小误差以增强鲁棒性。SVR 利用核方法将数据映射到高维特征空间，在高维空间中实现线性拟合，常用核函数包括高斯核、线性核和多项式核，以适应不同的非线性分布，其优化目标通过控制正则化参数 C 平衡模型复杂度与误差，同时仅依赖支持向量进行预测，具备良好的泛化能力，然而 SVR 对超参数（如 C、ϵ、核函数参数等）敏感，需要通过调优以获得最佳性能，适用于高维、非线性和噪声数据场景的回归任务。

2.8 极限梯度提升

XGBoost 是一种基于梯度提升框架的改进算法，通过引入系统优化和正则化技术，提升了模型的训练速度、预测精度以及对过拟合的控制能力。在回归任务中，XGBoost 通过优化目标函数并逐步拟合残差来逼近目标变量的真实值，其目标是构建一个由多棵回归树组成的模型，同时通过正则化技术控制模型复杂度以提升泛化能力。

为了优化目标函数，XGBoost 使用二阶泰勒展开，将损失函数近似为：

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N}\left[g_i f_{k,t}(x_i) + \frac{1}{2}h_i f_{k,t}^2(x_i)\right] + \Omega(f_{k,t}) \tag{2.16}$$

其中 g_i 和 h_i 分别是一阶梯度和二阶梯度，$\Omega(f_{k,t})$是正则化项，用于控制模型复杂度，XGBoost 每次分裂选择增益最大的分裂点，分裂增益计算公式为：

$$Gain = \frac{1}{2}\left[\frac{(G_L)^2}{H_L + \lambda} + \frac{(G_R)^2}{H_R + \lambda} - \frac{(G_{total})^2}{H_{total} + \lambda}\right] - \gamma \tag{2.17}$$

其中 G_L，G_R 是左右子节点的一阶梯度和；H_L，H_R 是左右子节点的二阶梯度和；γ 和 λ 是正则化参数。

2.9 评价指标

各种评价指标对评估机器学习模型是至关重要的，一些评价指标如下：

（1）均方误差（MSE）：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.18}$$

（2）均方根误差（RMSE）：

$$RMSE = \sqrt{MSE} \tag{2.19}$$

（3）平均绝对误差（MAE）：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.20}$$

（4）决定系数（R^2）：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.21}$$

（5）平均交叉验证得分为：

$$Mean\ CV\ Score = \frac{1}{K}\sum_{i=1}^{K} Accurancy_i \tag{2.22}$$

3 基于机器学习算法的实验过程

3.1 数据探索

3.1.1 特征变量分析

特征变量分析可以较全面了解数据的分布特性和基本结构，有助于发现数据质量问题，如缺失值、异常值或重复值，揭示变量间的相关性，为特征选择和特征工程提供依据，同时统计汇总能够评估模型对数据的适应性。通过统计汇总，还可以提高数据的可解释性，明确数据中潜在的规律，为后续的降维、聚类或建模奠定坚实基础，这一过程不仅是优化数据质量的必要环节，也是提升模型性能和决策效率的重要手段。

在该步骤中，本研究绘制了混凝土抗压强度 8 个特征变量的直方图和箱型图。

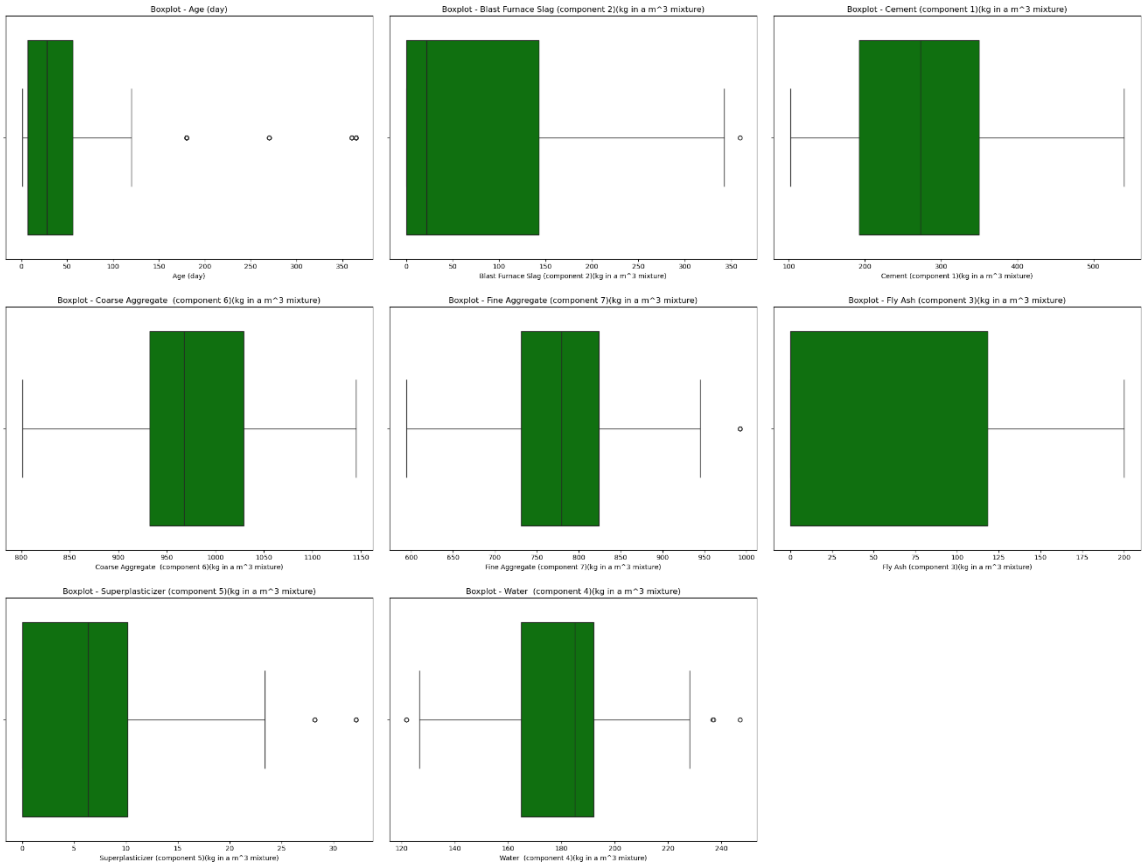

图 3-1 特征变量箱型图汇总

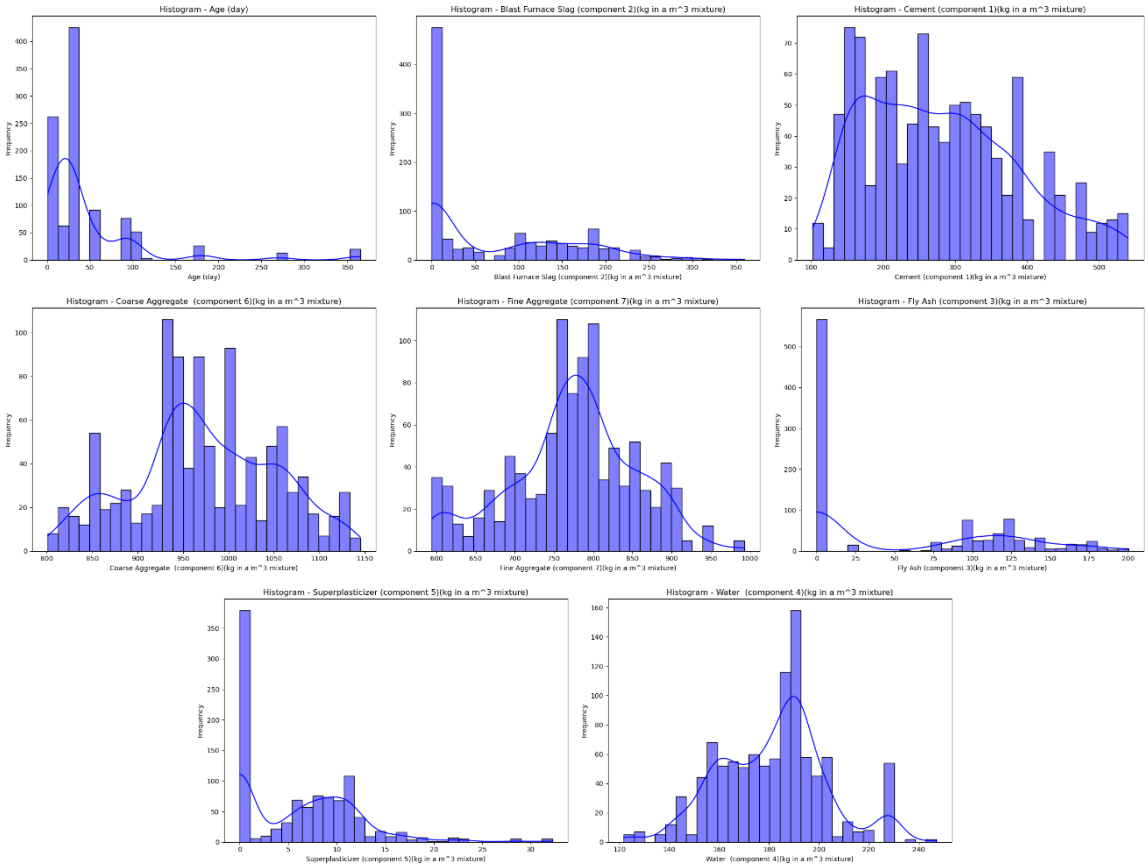

图 3-2　特征变量直方图汇总

3.1.2　目标变量分析

目标变量分析是数据探索中不可或缺的一部分，其核心在于全面理解预测目标的特性，为建模和数据处理提供坚实的基础。通过分析目标变量的分布形态，可以揭示集中趋势、离散程度以及潜在的异常值和缺失值，从而发现数据质量问题并采取相应的处理策略，例如分类任务中可以通过类别分布识别类别不平衡问题，而在回归任务中，偏态或多峰分布可能提示需要数据变换以提高模型效果，此外目标变量分析有助于评估任务复杂性，例如类别数量较多或类别边界模糊会增加分类难度，从而指导特定建模策略的选择。同时目标变量的特性还可以用于选择适当的模型评价指标，如类别不平衡时的宏平均 F1 分数或回归任务中的均方误差等。在特征工程方面，目标变量与特征变量的相关性分析可以帮助筛选关键特征，为模型构建提供方向。通过目标变量分析，不仅可以优化建模策略，还能提升数据的可解释性和预测的科学性，为最终模型的准确性和鲁棒性奠定基础。

混凝土抗压强度数据集的目标变量箱型图与直方图如图 3-3 所示。

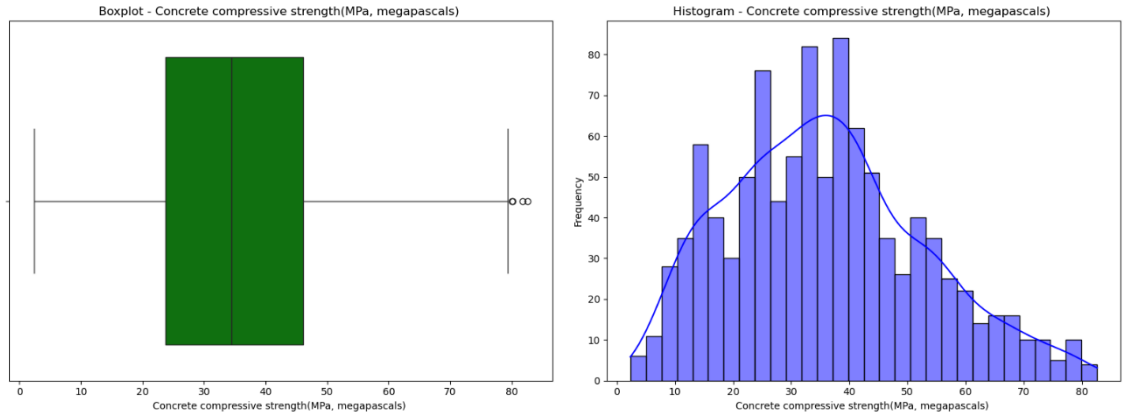

图 3-3　目标变量箱型图与直方图

目标变量混凝土抗压强度的分布特性显示其集中趋势和分散程度较为显著，根据箱线图和直方图分析，大多数样本的抗压强度集中在 30MPa 至 50MPa 之间，中位数约为 40MPa，呈现较为明显的集中性。然而分布具有轻微的正偏态，表现为部分高强度样本（大于 75MPa）的离群点，这些异常值可能代表特殊配方的混凝土。目标变量的取值范围较宽，从 5MPa 到 80MPa，数据整体呈单峰分布，且在低值和高值区间逐渐减少，反映了大部分样本的抗压强度集中于中等水平，因此建模时需关注离群点和正偏态可能带来的影响，需对目标变量进行适当的标准化或变换，以提高模型的拟合能力和预测性能。

3.1.3　变量相关性分析

变量相关性分析是数据探索中的重要步骤，用于揭示特征变量之间以及特征变量与目标变量之间的关系，为特征选择和模型设计提供科学依据。通过相关性分析，可以识别对目标变量具有显著预测力的关键特征，并剔除冗余或无关的特征，从而简化模型，提升训练效率和预测性能，例如计算皮尔逊相关系数或斯皮尔曼相关系数可以量化特征与目标变量的相关性，筛选出重要特征，同时通过分析特征间的强相关性，避免信息冗余和过拟合，此外相关性分析还能增强模型的解释性，帮助理解哪些特征对预测任务最重要，为业务优化或决策提供有力支持，因此相关性分析在提升模型性能和增强数据分析的科学性方面具有重要意义。

本研究中的混凝土抗压强度数据集的变量热力图如图 3-4 所示。

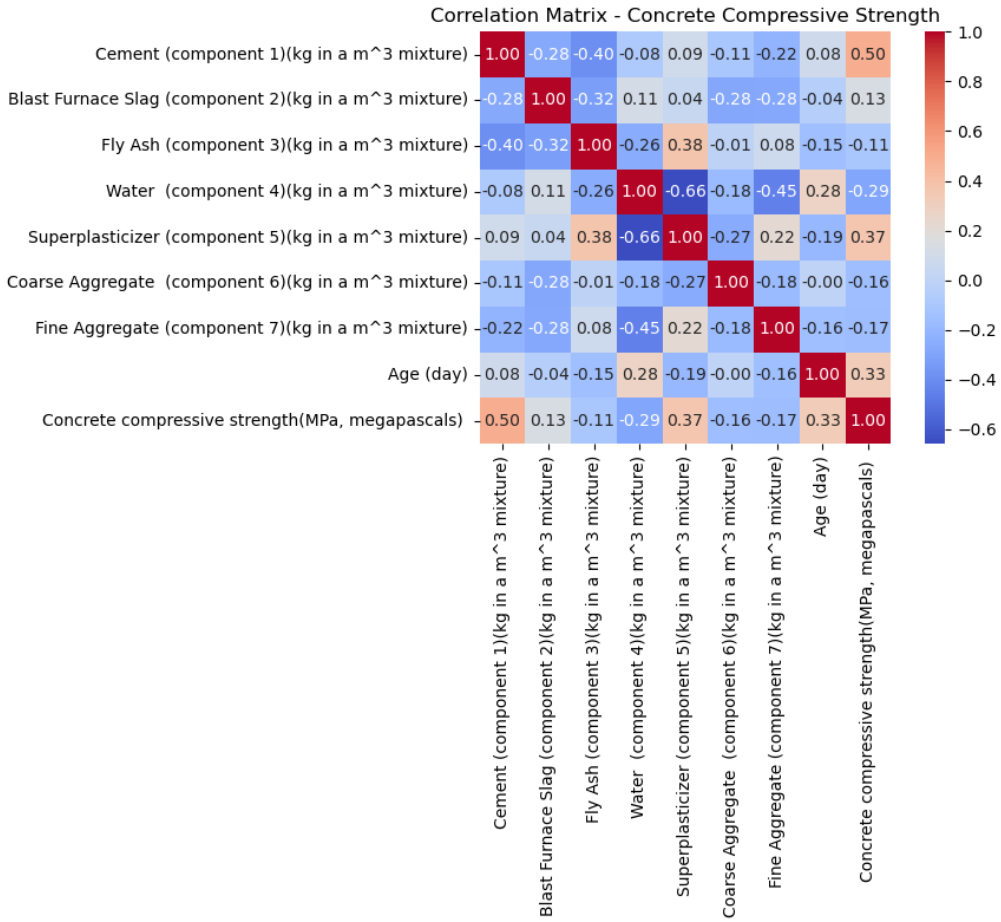

图 3-4　变量热力图

热力图揭示了混凝土抗压强度与多种输入变量的相关性特征，为构建预测模型提供了重要依据。水泥和龄期是影响抗压强度的主要正相关因素，相关系数分别为 0.5 和 0.33，表明它们对目标变量的预测能力较强，应作为关键输入特征纳入回归模型。粉煤灰和水与抗压强度分别呈中等和弱负相关（-0.40 和-0.29），这些特征可能对模型的拟合产生复杂的非线性贡献，在建模时需充分考虑其影响。其他特征如超塑化剂和骨料对目标变量的直接相关性较弱，但可能通过交互作用或非线性关系间接影响预测结果。结合这些相关性信息，回归任务可以利用多种方法充分挖掘主要特征对目标变量的影响，同时通过特征选择或交互项建模提升预测精度，热力图提供的相关性信息不仅有助于理解数据特征，还为构建高效、精准的回归模型奠定了基础。

3.1.4 异常值检测

本研究采用 IQR 方法进行异常值检测并移除，先计算四分位数：

$$Q1 = \text{quantile}(0.25) \tag{3.1}$$

$$Q2 = \text{quantile}(0.75) \tag{3.2}$$

由（3.1）和（3.2）式可得四分位距 IQR：

$$IQR = Q3 - Q1 \tag{3.3}$$

接着使用 1.5 倍的 IQR 来定义异常值的上下界，这个倍数是统计学中常用的经验值，然后过滤数据，令保留的数据满足以下（3.4）式子的条件。

$$Q1 - 1.5 \times IQR \leq x \leq Q3 + 1.5 \times IQR \tag{3.4}$$

3.2 数据预处理

3.2.1 数据清洗

由于原数据集作者声明没有缺失值，且经检查无误，于是按上文提到的 IQR 方法进行异常值检测，并清理变量异常值，具体效果如表 3-1 所示。

表 3-1 移除特征变量异常值数目表

特征变量	移除异常值数目
Cement	0
Blast Furnace Slag	2
Fly Ash	0
Water	9
Superplasticizer	10
Coarse Aggregate	0
Fine Aggregate	35
Age	44
Concrete compressive strength	4

3.2.2 特征工程

特征工程是从原始数据中提取描述性强的特征以优化模型性能的过程。这一过程非常关键，因为优秀的特征可以显著降低模型复杂性和参数调整的需要，使得简单模型也能达到高效的性能，所以好的特征不仅提升模型的表现，还增强了模型的灵活性

和维护的便捷性。

本研究的整个特征工程分为特征选择和标准化两个部分，根据特征热力图，选择跟目标标量相关性分数大于等于 0.1 的变量保留，并按照以下公式对混凝土抗压强度数据集进行标准化。

$$z = \frac{x - \mu}{\sigma} \tag{3.5}$$

其中 μ 为均值，σ 为方差。

3.3 模型训练和调参

3.3.1 规则树-线性回归

实验结果显示，规则树-线性回归模型在回归任务中表现出良好的预测能力。模型的 MSE 为 0.1305，RMSE 为 0.3612，MAE 为 0.2687，均表明预测误差较低，且误差幅度适中，R^2=0.8839 表明模型能够解释 88.39%的目标变量方差，拟合效果良好，MAE 小于 RMSE，说明大多数预测值与真实值的误差较小，少数较大的误差对整体性能影响有限。总体而言，该模型通过结合规则树的特征提取能力与线性回归的精准预测能力，在捕捉混凝土抗压强度的特征模式上取得了良好的平衡，适合当前数据特性，实验结果如表 3-2 所示。

表 3-2　规则树-线性回归模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
规则树-线性回归	0.1198	0.1305	0.3612	0.2687	0.8839

3.3.2 深度随机森林

实验结果显示，深度随机森林模型在回归任务中表现出较强的非线性拟合能力。模型的 MSE 为 0.1635，RMSE 为 0.4043，MAE 为 0.2930，均表明预测误差稍高于规则树-线性回归模型，误差幅度相对较大，R^2=0.8545 表明模型能够解释 85.45%的目标变量方差，拟合效果较为良好但略逊于规则树-线性回归模型。MAE 小于 RMSE，说明大多数预测值与真实值的误差较小，但少数较大的误差对整体性能的影响稍显显著。总体而言，深度随机森林模型通过其强大的非线性建模能力，在处理复杂特征分布时

展现出一定的优势，但对混凝土抗压强度的预测仍有优化空间，实验结果如表 3-3 所示。

表 3-3　深度随机森林模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
深度随机森林	0.1600	0.1635	0.4043	0.2930	0.8545

3.3.3　极端随机树

实验结果显示，极端随机树模型在回归任务中展现了优异的性能。模型的 MSE 为 0.1024，RMSE 为 0.3200，MAE 为 0.2146，均表明预测误差较低，且误差幅度相对较小，R^2=0.9089 表明模型能够解释 90.89%的目标变量方差，拟合效果优于规则树-线性回归模型和深度随机森林模型。MAE 小于 RMSE，说明大多数预测值与真实值的误差较小，少数较大的误差对整体性能影响有限，说明极端随机树模型通过其高效的特征分裂策略和集成学习能力，在捕捉混凝土抗压强度特征模式上表现出色，是当前任务中预测能力最强的模型之一，实验结果如表 3-4 所示。

表 3-4　极端随机树模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
极端随机树	0.0920	0.1024	0.3200	0.2146	0.9089

3.3.4　梯度提升决策树

实验结果显示，梯度提升决策树模型在回归任务中表现良好，但相较于极端随机树模型略逊一筹。模型的 MSE 为 0.1210，RMSE 为 0.3478，MAE 为 0.2537，均表明预测误差较低且误差幅度适中，R^2=0.8923 表明模型能够解释 89.23%的目标变量方差，拟合效果优于深度随机森林模型，但略低于规则树-线性回归模型和极端随机树模型。MAE 小于 RMSE，说明大多数预测值与真实值的误差较小，少数较大的误差对整体性能的影响有限。总体而言，梯度提升决策树模型通过其梯度优化机制展现了良好的特征学习能力和泛化能力，是一种在当前任务中具有竞争力的回归模型，实验结果如

表 3-5 所示。

表 3-5　梯度提升决策树模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
梯度提升决策树	0.0981	0.1210	0.3478	0.2537	0.8923

3.3.5　轻量级梯度提升机

实验结果显示，轻量级梯度提升机在回归任务中展现出优异的性能。模型的 MSE 为 0.0871，RMSE 为 0.2952，MAE 为 0.2071，均表明预测误差较低，且误差幅度小于其他模型，R^2=0.9225 表明模型能够解释 92.25%的目标变量方差，拟合效果优于规则树-线性回归模型、深度随机森林模型、极端随机树模型和梯度提升决策树模型，MAE 小于 RMSE 说明大多数预测值与真实值的误差较小，少数较大的误差对整体性能的影响极小。总体而言轻量级梯度提升机通过其高效的特征处理和优化机制，在当前任务中取得了最优的预测效果，是最具竞争力的回归模型之一，实验结果如表 3-6 所示。

表 3-6　轻量级梯度提升机模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
轻量级梯度提升机	0.0836	0.0871	0.2952	0.2071	0.9225

3.3.6　随机森林

实验结果显示，随机森林模型在回归任务中表现出了良好的预测能力。模型的 MSE 为 0.1064，RMSE 为 0.3262，MAE 为 0.2323，表明预测误差较低，且误差幅度适中，R^2=0.9053 表明模型能够解释 90.53%的目标变量方差，拟合效果优于深度随机森林模型和梯度提升决策树模型，但略逊于轻量级梯度提升机和极端随机树模型，MAE 小于 RMSE，说明大多数预测值与真实值的误差较小，少数较大的误差对整体性能的影响较小。总体而言，随机森林模型通过其强大的集成学习能力，在处理复杂数据特征方面展现了较强的鲁棒性和适应性，是当前任务中表现较为稳定的回归模型之一，

实验结果如表 3-7 所示。

表 3-7　随机森林模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
随机森林	0.1095	0.1064	0.3262	0.2323	0.9053

3.3.7　支持向量机

实验结果显示，支持向量机模型在回归任务中的表现相对较为一般。模型的 MSE 为 0.1421，RMSE 为 0.3769，MAE 为 0.2708，均表明预测误差略高于随机森林、轻量级梯度提升机等模型。R^2=0.8736 表明模型能够解释 87.36%的目标变量方差，拟合效果相较其他模型略显不足，MAE 小于 RMSE 说明大多数预测值与真实值的误差相对较小，但少数较大的误差对整体性能造成了一定影响。支持向量机模型在捕捉混凝土抗压强度的特征模式方面具备一定能力，但其在处理复杂非线性关系和大规模数据时的表现相对有限，实验结果如表 3-8 所示。

表 3-8　支持向量机模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
支持向量机	0.1343	0.1421	0.3769	0.2708	0.8736

3.3.8　极限梯度提升

实验结果显示，极限梯度提升模型在回归任务中展现了极高的预测精度。模型的 MSE 为 0.0903，RMSE 为 0.3006，MAE 为 0.2000，均表明预测误差较低且误差幅度最小，优于除轻量级梯度提升机外的大多数模型，R^2=0.9196 表明模型能够解释 91.96%的目标变量方差，拟合效果极佳，接近轻量级梯度提升机的性能，MAE 小于 RMSE 说明大多数预测值与真实值的误差较小，少数较大的误差对整体性能影响有限，总体而言，极限梯度提升模型凭借其强大的优化机制和非线性特征建模能力，在当前任务中表现卓越，是一种高效且精准的回归模型选择之一，实验结果如表 3-9 所示。

表 3-9　极限梯度提升模型实验结果

模型	平均交叉验证得分	均方损失	均方根损失	平均绝对误差	决定系数
极限梯度提升	0.0784	0.0903	0.3006	0.2000	0.9196

3.4　模型对比分析

各模型的 RMSE 结果汇总如表 3-10 所示，展示了不同回归模型在预测混凝土抗压强度任务中的性能差异。从结果来看，LightGBM 模型以 RMSE=0.2952 的表现最优，预测误差最小，显示其在处理大规模数据和非线性关系方面的强大能力，其次是 XGBoost 和 Etr 模型，分别达到 RMSE=0.3006 和 RMSE=0.3200，表现出接近的高预测精度，随机森林和 GBDT 模型的 RMSE 分别为 0.3262 和 0.3478，略逊于上述模型，但仍保持较高的预测能力。相比之下 Cubist 模型和深度随机森林模型的 RMSE 分别为 0.3612 和 0.4043，误差相对较大，预测精度较低，而 SVR 模型的 RMSE 为 0.3769，性能介于 Cubist 与 GBDT 之间。

综上所述 LightGBM 模型凭借最低的 RMSE 在本实验中表现最佳，适合作为预测混凝土抗压强度的首选模型，而 XGBoost 和 Etr 模型表现同样优异，具有一定竞争力。其他模型在性能上略逊一筹，但依然展现了较强的回归能力，能够作为不同场景下的备选方案。

表 3-10　模型均方根损失对比

模型	均方根损失
Cubist	0.3612
Deep RF	0.4043
Etr	0.3200
GBDT	0.3478
LightGBM	0.2952
RF	0.3262
SVR	0.3769
XGBoost	0.3006

3.5 本章小结

本章以混凝土抗压强度数据集为研究对象，全面探讨了基于机器学习算法的回归实验过程。首先通过数据探索对特征变量和目标变量进行了分析，总结了数据的分布特性、变量相关性和潜在的异常值，为后续建模奠定了坚实基础。变量相关性分析表明，水泥含量和龄期是与抗压强度关系最为密切的关键变量，而粉煤灰和水的负相关性提示了可能的非线性影响，为特征选择提供了科学依据。

在数据预处理阶段，本研究采用了 IQR 方法检测并移除异常值，并对保留的特征进行了标准化处理，从而有效提升了模型训练的稳定性和预测的准确性。随后构建并评估了多种回归模型，包括规则树-线性回归、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机和极限梯度提升。通过对模型的均方损失、均方根损失、平均绝对误差和决定系数等指标的对比分析，全面评估了各模型在预测混凝土抗压强度任务中的性能表现。

实验结果显示轻量级梯度提升机和极限梯度提升模型在预测精度和泛化能力方面表现最佳，分别以最低的 RMSE（0.2952 和 0.3006）和最高的 R^2（0.9225 和 0.9196）领先其他模型，极端随机树和随机森林模型表现亦较为优异，均在处理非线性关系和特征交互方面展现出强大能力。相较之下深度随机森林和 Cubist 模型的预测误差相对较大，而支持向量机模型在复杂非线性关系建模中的表现稍显不足。

综上所述，本章通过对不同机器学习算法的系统性比较，验证了轻量级梯度提升机和极限梯度提升在回归任务中的卓越性能，尤其适合混凝土抗压强度的预测任务，同时实验揭示了数据特性、特征选择及模型优化在提升模型性能中的重要作用，为后续研究提供了科学指导和实践参考。

4 总结与展望

4.1 全文总结

本研究以混凝土抗压强度数据集为基础，系统评估了多种机器学习模型在回归任务中的性能表现。通过对规则树-线性回归、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机和极限梯度提升等模型的实验比较，结果显示轻量级梯度提升机和极限梯度提升以最低的均方根误差（RMSE=0.2952 和 0.3006）和最高的决定系数（R^2=0.9225 和 0.9196）表现最佳，极端随机树和随机森林模型也展现了优异的预测能力和鲁棒性，本研究通过数据预处理和特征工程优化了模型输入特征，提高了模型的稳定性和准确性，同时揭示了集成学习模型在捕捉非线性关系和特征交互方面的显著优势。研究结果为混凝土抗压强度的精准预测提供了科学支持，并为实际工程应用中的模型优化和算法改进指明了方向。

4.2 展望

由于时间原因，本研究中大量模型没有进行细致的调参，没有将精度提升到这个模型应有的水平，另外本研究中还有大量的实验数据和评价指标做出了但没有进行可视化对比，以后时间充裕时，会在 github 仓库中进行可视化图表以及复现步骤的补充。

致 谢

在本次课程的学习过程中，我深刻体会到了机器学习的魅力与挑战。这门课程不仅拓宽了我的知识视野，还培养了我在数据处理与模型分析方面的实践能力。在此衷心感谢老师的悉心指导和无私分享，严谨的教学态度和丰富的专业知识让我受益匪浅。除此之外，也感谢课程为我提供了利用高质量教学资源和进行实验的机会，使我能够将理论知识有效应用于实际问题。这段学习经历将成为我未来研究与实践的宝贵财富，非常感谢！机器学习课程自开课以来，就需要一代代课程团队不断打磨，希望咱们学校的机器学习课程能够越办越好！