
# 南京大学

# 本科毕业论文

院　　系　计算机科学与技术系

专　　业　计算机科学与技术

题　　目　基于多变量回归分析的健康成年人脑龄预测研究

年　　级　2018　学　　号　185220006

学生姓名　郭珉铉

指导教师　高阳　职　　称　教授

提交日期　2024 年 6 月 6 日

# 南京大学本科毕业论文（设计）
# 诚信承诺书

本人郑重承诺：所呈交的毕业论文（设计）（题目：基于多变量回归分析的健康成年人脑龄预测研究）是在指导教师的指导下严格按照学校和院系有关规定由本人独立完成的。本毕业论文（设计）中引用他人观点及参考资源的内容均已标注引用，如出现侵犯他人知识产权的行为，由本人承担相应法律责任。本人承诺不存在抄袭、伪造、篡改、代写、买卖毕业论文（设计）等违纪行为。

作者签名：郭子强

学号：185220006

日期：2024年6月5日

# 南京大学本科生毕业论文（设计、作品）中文摘要

**题目**：基于多变量回归分析的健康成年人脑龄预测研究

**院系**：计算机科学与技术系

**专业**：计算机科学与技术

**本科生姓名**：郭珉铉

**指导教师（姓名、职称）**：高阳　教授

**摘要**：

脑龄预测对于识别出标准大脑老化过程中的异常至关重要，这些异常可能是神经退行性疾病的早期迹象。如果脑龄显著高于实际年龄，这可能表明大脑加速老化，是神经退行性疾病的一个潜在早期标志。本研究基于 MRI 成像数据衍生的包含 410 个特征的高维数据集，探索了多种特征工程技术对预测模型性能的提升效果。首先，本研究采用了综合的特征创建过程和严格的特征选择策略，特别是引入了分组聚合特征，以强化数据集的信息量和模型的预测能力。其次，通过比较不同特征工程技术对各种预测模型的影响，确保了方法论的全面性和科学性。然后，研究结果显示，应用这些特征工程技术后，LightGBM 模型在所有比较中表现最佳，突出了其在处理高维数据时的优越性。另外，分组聚合特征的引入最为有效地提高了模型性能。结合两者得到了本文最终脑龄预测模型。本研究突显了特征工程的重要性，适当且有效地施行特征工程技术可提升预测精度和模型鲁棒性。这对于神经退行性疾病的早期检测和干预具有重要意义。

**关键词**：机器学习；回归；脑龄预测

# 南京大学本科生毕业论文（设计、作品）英文摘要

THESIS: Predicting Brain Age in Healthy Adults Using Multivariate Regression Analysis

DEPARTMENT: Department of Computer Science and Technology

SPECIALIZATION: Computer Science and Technology

UNDERGRADUATE: Koay Min Sen

MENTOR: Gao Yang，Professor

ABSTRACT:

Brain age prediction serves as a crucial marker in identifying deviations from typical aging patterns, potentially indicating neurodegenerative diseases. This study explores the efficacy of various feature engineering techniques in enhancing the predictive accuracy of models trained on a high-dimensional dataset derived from MRI imaging data, comprising 410 original features. Our approach included a comprehensive feature creation procedure, where grouped aggregate features were introduced, alongside rigorous feature selection methods. We assessed the performance impact of these engineering techniques on several predictive models, concluding that Light Gradient Boosting Machine (LightGBM) outperforms other models. The implementation of grouped aggregate features notably improved model performance, establishing them as the most effective feature engineering technique in this study. Our findings underscore the importance of sophisticated feature engineering in developing robust predictive models for brain age, which could be integral in early detection and intervention strategies for neurodegenerative conditions.

KEYWORDS: Machine Learning; Regression; Brain Age Prediction

# 目　录

# 第一章　导论

## 1.1　研究背景

人类大脑，作为中枢神经系统的核心组成部分，承载着复杂的思维、情感、记忆以及对身体运动的高度协调功能。其独特性不仅体现在其构造的复杂性上，更在于其在人类行为和心理活动中的核心作用[1]。大脑由数十亿个神经元以及更多的胶质细胞组成，这些细胞通过一个复杂的电化学信号网络相互交流，形成了思考、感知、决策和学习的生物基础。显然，大脑是人类最重要器官之一。

脑龄是指人类大脑生理状态或功能状态相对于其实际年龄的估计年龄。这个概念通常用于衡量大脑健康和老化的速度，可以通过各种神经影像学方法和认知测试来评估。脑龄的计算通常基于大脑结构的变化（如皮层厚度、脑体积等）或功能的变化（如记忆力、注意力等）。基于结构磁共振的大脑年龄被广泛应用于刻画大脑的老化过程，预测脑龄和实际生理年龄的差值（Predicted Age Difference，PAD），即偏离正常大脑老化轨迹的程度，可作为衡量个体异常老化的客观指标。研究表明多种类型的神经系统疾病、代谢性疾病等都与大脑异常老化相关[2]。

脑龄预测传统方法结合了神经学、放射学和心理学等多个领域的专业知识，侧重于分析与老化相关的大脑结构和功能变化以及认知测试。传统的神经影像学技术，如 MRI（磁共振成像）和 CT（计算机断层扫描），被广泛用于观察和测量大脑中的物理变化[3]。专家们会亲自评估大脑体积、皮层厚度、白质完整性和脑室扩大等老化指标的变化再与已知的与年龄相关的变化进行比较。认知测试也属于一种预测脑龄的传统方法之一。该测试可以在没有计算机辅助的情况下进行，结果将与不同年龄组的标准数据进行比较，以评估个体的认知表现是否与其实际年龄一致。传统方法面临着诸多挑战，包括分析过程耗时、精确度有限以及高度依赖专家知识和主观判断等问题。

随着人工智能和机器学习技术的迅速发展，脑龄预测任务迎来了重大进步。

机器学习模型，已经被开发和应用于自动化分析大量的神经影像数据，这些算法能够从复杂的数据中学习到年龄相关的生物标志物和模式[4]。相较于传统方法，机器学习技术的引入不仅大幅提高了脑龄预测的效率和精确度，而且还减少了对专家主观判断的依赖，使得预测过程更加客观和可重复。此外，机器学习模型的高度可扩展性和灵活性使得它们能够跨不同人群和数据集进行有效的脑龄预测，从而为早期诊断神经退行性疾病、监测治疗效果以及个性化医疗提供了有效的工具。

## 1.2 研究意义

现代社会正面临着人口老龄化的挑战，而脑龄预测作为一种相对新兴的研究领域，对于维护和提升成年人的脑健康水平具有重要意义。本研究旨在探索成年人脑龄预测的重要性与意义，从健康管理、疾病预防、神经科学研究以及生活方式干预等方面进行分析和论述。

首先，成年人脑龄预测对于健康管理和个体关怀具有重要意义。通过预测成年人的脑龄，可以帮助个体了解自身脑部健康状况，及早发现潜在问题并采取相应措施。这有助于提升个体的健康意识，促进自我管理和健康行为的形成，从而降低患神经退行性疾病的风险，提高生活质量。可靠的脑龄预测也可以为个体提供生活方式干预和个性化健康管理提供科学依据。我们可以基于脑龄预测结果制定针对个体的生活方式干预方案，如调整饮食、增加运动、改善睡眠质量等，从而延缓脑龄的加速老化，提高健康水平，促进个体全面发展。

其次，生物指标是用来量化生物体状态或特性的指标，而预测脑龄是通过各种生物学和神经影像学特征对个体的脑结构和功能进行估计，因此可以被归类为生物指标的一种。预测脑龄的结果可以反映个体的脑健康状况，对于研究神经系统的发育、老化、疾病以及与之相关的因素具有重要的生物学意义。随着人口老龄化趋势的加剧，神经退行性疾病的发病率逐渐增加，给社会和个体带来了巨大负担。使用 MRI 等神经影像技术进行的大脑年龄预测，通过将个体预测的大脑年龄与其实际年龄进行比较, 大脑年龄显著老于实际年龄的时候，一般穷况下表明神经退化的早期迹象。因此通过脑龄预测技术，可以及早发现与神经退行性疾病相关的早期迹象，为个体提供预防和干预的机会，有望减缓或阻止疾病的

发展，降低医疗资源的压力。Ran 等人[5]提出的结合回归技术与 Shapley Additive Explanation(SHAP) 的大脑年龄向量指标，在早期疾病筛查中表现优于其他大脑老化指标和大脑体积，证明了其在早期检测神经退行性疾病方面的有效性。

最后，成年人脑龄预测有助于促进神经科学研究的发展。了解成年人脑龄与脑功能之间的关系，对于深入探究脑部结构和功能的变化规律，揭示脑衰老的机制具有重要意义。这不仅有助于解决基础科学问题，还为开发针对性的脑健康干预手段提供理论支持。Sihag 等人[6]提出了使用协方差神经网络（VNN）利用皮层厚度特征创建解剖学上可解释的大脑年龄预测的方法，为阿尔茨海默病中导致大脑年龄增高的区域提供了见解。

综上所述，成年人脑龄预测的研究意义在于促进健康管理、疾病预防、神经科学研究以及生活方式干预等方面都具有重要价值。通过深入探索成年人脑龄预测的意义，有助于更好地理解和应用脑龄预测技术，为维护和提升成年人的脑健康水平提供科学依据和理论支持。

## 1.3 本文工作

本文将来自全国 15 所大学医院应用 1.5 T MR 共采集 3000 名正常成年志愿者的全脑结构磁共振图像。数据集经过严格质控后的健康成年人近 1600 例，受试者资料包括性别、年龄、体积、表面积、厚度、平均曲率、高斯曲率以及 MRI 扫描仪类型。该数据集的排除标准为患有任何神经精神疾病和磁共振数据质量。本文将结合相关知识，构建必要的特征工程，建立机器学习、人工智能或数据挖掘模型，并用该模型预测成年人脑龄。

本文核心技术在于对年龄段进行分段并计算各段的统计计量（如中位数、均值和标准偏差等），这一方法显著提高了模型捕捉复杂信息的能力。通过这种创新技术，本文的模型不仅能更准确地预测成年人的脑龄，还能自动化地处理大量数据，显著减少了传统脑龄预测过程中的时间消耗和对专家知识的依赖。传统的脑龄预测方法不仅耗时而且精确度有限，且高度依赖专家的主观判断。通过引入机器学习技术，我们的模型能够自动识别与脑龄相关的复杂模式，从而解决了以往方法的局限性。

此外，这种技术的引入对医学领域的贡献尤为显著。它为早期诊断神经退行

性疾病如阿尔茨海默病提供了一个强有力的工具，可以通过分析脑结构的微小变化来预测疾病的发展。该模型的高效性和精确性也使得它可以广泛应用于临床实践和医学研究，帮助医生和研究人员更好地理解脑老化过程及其对健康的影响。

## 1.4 论文结构

本文一共分为五个章节：

第一章为导论。首先介绍了本文研究背景、研究意义以及本文主要研究内容，强调了脑年龄预测在医学和科学研究中的应用重要性。此外，最后介绍了本文的总体结构。

第二章为文献综述。本章深入探讨了传统脑年龄预测技术面临的主要挑战，并分析了机器学习技术在解决这些问题中的潜力。本章还介绍了机器学习技术的常见研究方法，并简单介绍了本文所采用的机器学习模型与特征工程技术在高维数据的应用。

第三章为模型构建。本章首先简单介绍了数据集的与处理方式。之后，主要介绍了本文对高维数据的处理手段，叙述了特征工程的构建思路。接着介绍了本文所使用的机器学习模型与具体实现方法。最后介绍了本文所使用的模型性能评价指标。

第四章为实验结果与分析部分。首先介绍了本文的实验流程。接着记录具体各个机器学习模型在有效的特征工程技术下所得到的模型性能提升，并使用一致的模型性能评价指标进行分析和结果对比。

第五章为结论和展望，对本文进行概述，总结了所使用的特征工程技术对于预测模型的实验结果，给出了本文主要研究结论。最后讨论了后期工作展望和工作的可改进方向。

# 第二章　文献综述

## 2.1　脑龄预测的传统方法

传统大脑年龄评估方法尝试通过直接观察这些变化来推断个体的神经生物学健康状况。从神经影像学的精细扫描到复杂的认知功能测试，每种技术都旨在从不同的角度解读大脑老化的特征。这些方法不仅有助于识别与年龄相关的正常变化，还有助于早期发现潜在的神经退行性疾病。

MRI 和 CT 扫描这样的神经影像技术在评估大脑结构和功能方面起到了关键作用。Raz 等人[7]的研究广泛记录了大脑萎缩和白质完整性丧失的模式，以上现象都是大脑老化的指标。这些结构变化与认知下降相关，对于早期检测神经退行性疾病至关重要。

认知测试是评估大脑健康的另一个基石。Bernard[8]等人提出了简易智力状态检查（MMSE），旨在对认知功能进行简要的定量评估，一般适用于识别认知功能障碍和痴呆。MMSE 的结构包括各种认知领域，包括定向、记忆、注意力、语言和视觉构建能力。Nasreddine[9]等人的蒙特利尔认知评估（MoCA），是用于检测轻度认知功能障碍（MCI）的工具。MoCA 在检测 MCI 方面表现出很高的灵敏度（90）和特异度（87），使用 26 分的截断分数。该研究表明 MoCA 在早期检测认知衰退相比于 MMSE 更胜一筹。[10][11]各提出了 MoCA 的改编版本，分别为巴西和韩国版，也经过验证，表现出良好的可靠性和在不同人群中识别 MCI 的一致性表现。这些改编版本证实了 MoCA 在不同文化和语言环境中的广泛适用性和有效性。

除此之外，Jack 等人[12]讨论了使用脑脊液中的淀粉样 β 蛋白和 tau 蛋白等生物标志物来预测阿尔茨海默病的发病。这些标志物提供了关于大脑健康的生化洞察，能在临床症状表现之前指示神经退行性变化。Babiloni 等人[13]（2010）强调了 EEG 模式的变化，特别是前额皮层的变化，与认知下降有关。这些方法对于持续监测很有用，可以检测出通过影像或认知测试单独可能无法见到的变化。

Petersen[14]详细说明了使用临床标准来诊断轻度认知障碍等病症，这是更严重认知障碍的前兆。这种评估对于制定治疗计划和管理认知衰退的进程至关重要。

## 2.2 机器学习技术在脑龄预测的应用

数据的广泛可用性为机器学习提供了必要的基础与研究对象，计算资源的不断提升加快了模型训练和迭代过程，高效算法对模型的优化大幅度提高了机器学习在各个领域中的应用效果。这些因素共同促成了机器学习技术的快速发展和广泛应用。机器学习技术在大脑老化领域也获得不少关注。

相对与脑龄预测的传统方法，机器学习技术能够分析复杂的，高维度的数据。包括但不仅限于大量的影像数据、由影响数据派生出来的表格数据等等。能够检测到人类观察者可能察觉不到的大脑结构和功能的微妙变化。这种能力导致更准确的大脑年龄预测和对异常大脑老化或神经退行性疾病的早期检测。Frank等人[2]介绍了 BrainAGE 框架，一种基于结构性 MRI 数据的新颖方法，可可靠地估计大脑年龄，并揭示了不同诊断组之间大脑衰老差异的内在规律。

机器学习技术相比于传统方法的不同之处也体现于可对多种数据类型整合到一个预测模型中，因此得到一个更为全面的预测模型。Kramer 等人[15] 探讨了利用多模态神经影像数据（包括区域灰质体积（GMV）、静息态功能连接（RSFC）和结构连接（SC））来预测成年人认知表现。

机器学习技术，特别是深度学习，可以从复杂的数据集中提取未被人类研究者显式编程或预期的特征。在神经影像学中，深度学习模型可以识别与老化或疾病相关的独特模式，这些模式超出了典型的如体积或密度等测量。Cole 等人[16]该研究使用卷积神经网络（CNN）在一个包含 2001 名健康成年人的大型数据集上，演示了准确预测大脑年龄的方法。这种方法能够准确预测出生物年龄，这对于理解大脑衰老的偏差至关重要。Jonsson 等人[17]研究提出了一种新颖的深度学习方法，利用残差卷积神经网络从 T1 加权 MRI、雅可比图以及灰质和白质分割图像中预测大脑年龄，以研究年龄相关的结构性大脑变化与生物年龄之间的差异。

机器学习方法的主要优势之一是在各种任务中显著降低人力投入，其中脑龄预测亦不例外。预测模型一旦训练完成，可以迅速且大规模地评估新的案例。

此外，传统脑龄预测方法特别是认知测试和神经影像学中存在的一个关键问题是不同从业者间的观察者变异，而机器学习技术的数据驱动特性有效地缓解了这一问题。同时，应用机器学习技术还可以消除评估者的主观性，提高预测的客观性。Tanveer 等人[18] 综述讨论了深度学习模型如何提供更加标准化的年龄预测方法，相比传统的放射学评估，减少了主观性，这种主观性在观察者之间可能存在显著差异。

机器学习在大脑年龄预测方面相比传统方法具有显著的进步，它通过提高准确性、效率和分析深度来实现。与传统方法依赖于手动选择特征和简单统计模型不同，机器学习算法可以自动识别大数据集中的复杂模式和相关性，包括 MRI 图像和临床生物标志物。这种能力不仅提高了年龄预测的精度，还能够检测到传统方法可能忽略的神经退行性变化的微妙早期迹象。此外，机器学习模型，特别是深度学习技术，可以处理多模态数据源，整合各种类型的信息，以提供对大脑健康的更全面评估。这种整合有助于更深入地理解与衰老和疾病进展相关的潜在生物过程，为有针对性的干预提供潜在途径。最终，机器学习能够从数据中学习并适应新信息的能力，为大脑年龄预测和更广泛的神经学研究领域提供了可扩展且可能更有效的方法。

## 2.3 机器学习技术的主要研究方法

成年人脑龄预测是近年来神经科学和人工智能交叉领域的一个热门研究方向。它旨在通过分析大脑的成像数据来估计一个人的脑年龄，这可以作为评估个体大脑健康状况和认知功能的一个重要指标。传统的脑龄预测方法依赖于医学影像专家对大脑 MRI（磁共振成像）图像的分析，这一过程不仅耗时而且成本高昂，其准确性很大程度上取决于专家的经验和判断，这导致了结果的主观性和不一致性。机器学习技术成功提高了大脑年龄预测任务的精度。

### 2.3.1 线性回归

线性回归 (Linear Regression) 是一种常见的统计建模技术，用于分析自变量和因变量之间的线性关系。这是一种对线性方程的系数进行缩放的建模技术。

线性回归模型可以表示为以下方程：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon, \tag{2-1}$$

其中：$y$ 是我们试图预测或解释的因变量。$\beta_0$ 是截距项，表示当所有自变量 $x$ 为零时的 $y$ 值。$\beta_1, \beta_2, \cdots, \beta_n$ 是自变量 $x_1, x_2, \cdots, x_n$ 的系数。这些系数代表了在其他自变量保持不变的情况下，自变量变化一个单位时因变量的变化量。$\epsilon$ 是误差项，它解释了 $y$ 中无法通过线性模型解释的变异性。

在实际应用中，用于训练线性回归模型的数据往往都具有高维度和多重共线性的特点，该特点是线性回归模型面临的巨大挑战。为了应对这些挑战，研究人员提出了多种正则化方法，以改善模型的预测性能和泛化能力。Hoerl 等人[19]的论文主要贡献在于在最小二乘估计中引入了一项正则化项，有效地控制了模型的复杂性，并在处理相关或大规模的预测变量集时防止过拟合. Tibshirani 等人[20] 的研究具有开创性的工作奠定了回归方法的基础，不仅通过其将某些变量的系数设定为零来提供收缩，还通过变量选择来简化模型。[21]使用了以 L1 和 L2 正则项，结合了 Lasso 回归和 Ridge 回归的核心，在观测变量数量多于观测次数的情境，或者多个变量高度相关的情况下特别实用。

### 2.3.2 梯度提升决策树

梯度提升决策树（Gradient Boosting Decision Trees, GBDT），是一种集成学习方法。它通过反复迭代训练决策树模型来提升预测性能。GBDT 的基础模型是决策树，每个决策树都是一棵深度较浅的树，由一个节点和两个叶子节点组成。在训练过程中，GBDT 逐步构建一系列决策树，每一棵树都试图纠正前一棵树的残差 (residual)，尽可能把损失最小化。GBDT 最初由[22]提出。该研究介绍了梯度提升的统计框架和算法，将弱学习者结合起来形成一个强大的预测模型。GBDT 在机器学习中被广泛应用，尤其是在回归和分类问题中取得了显著的成功。

#### 2.3.2.1 XGBoost

XGBoost 是由[23]提出的 GBDT 框架。XGBoost 被描述为基于预排序的决策树算法因为它采用了一种特殊的排序算法来优化分裂点的选择。决策树在每一

层生长之前，首先必须进行特征值的排序。一旦完成排序，XGBoost 遍历这些已排好序的特征值寻找最优分裂点，最大化信息增益。这算法不仅需要保存数据的特征值，还得存储排序后的索引，导致空间上和时间上的昂贵开销。

传统的 GBDT 实现一般采用贪婪（Greedy）的决策树生长策略，这是一种开销庞大且非常占用内存方法。XGBoost 使用了按层（level-wise）决策树生长策略，是一种在每一步中同时对树的同一层所有节点进行分裂的方法。这意味着每一层的所有叶子节点，无论它们的分裂增益如何，都会在进入下一层之前尝试分裂。在硬件层面上，按层生长可以更高效地利用缓存，因为算法可以顺序地访问内存中相邻的数据。

此外，XGBoost 在损失函数中加入了正则化项（包括树的叶子节点的权重的 L1 和 L2 正则化），这不仅帮助在训练过程中控制模型的复杂度，还防止了过拟合。再者，XGBoost 引入的列抽样技术，不仅可提供了额外的降低过拟合风险的效果，还显著提高大规模数据集的处理速度。显然，XGBoost 也有不足之处。XGBoost 的按层 (level-wise) 生长策略, 即使分裂增益不大，还是会进行分裂，导致对计算资源的浪费。

#### 2.3.2.2 LightGBM

LightGBM (Light Gradient Boosting Machine) 是由微软 2016 年开发的 GBDT 开源框架[24]。它以其高效性和可扩展性而闻名，特别擅长处理大型和高维数据集。LightGBM 框架针对速度和内存使用进行了优化，使其比起其他基于树的学习算法更高效。LightGBM 在传统的 GBDT 算法上进行了如下优化：

首先是决策树生长策略的改进。LightGBM 使用的是带深度限制的 Leaf-wise 的叶子生长策略。Leaf-wise 方法通过选择使损失函数减少最大的叶子节点进行分割来生长树。这种有针对性的方法可以快速减少误差，尤其是在具有复杂模式和高维度的数据集中。其缺点在于有可能会生长出一个过于深的决策树，导致模型过拟合。为了保证模型保持高效和降低过拟合风险的同时，LightGBM 选择在 leaf-wise 之上增加最大深度限制。

其次，在高维数据中，许多特征可能是稀疏的，意味着它们主要包含零值。互斥特征捆绑算法（EFB）是 LightGBM 的一项具有创新性功能，它将互斥特征（很少同时取非零值的特征）捆绑到单个特征中。这减少了模型需要处理的特征

数量，降低了内存消耗，提高了训练速度的同时，尽可能保持模型的准确性。

基于梯度的单边采样（Gradient-based One-Side Sampling，GOSS）算法通过排除多部份信息增益不大地样本，而达到了降本增效的效果。换句话说，GOSS算法在进行数据采样的时候基本上只保留梯度大的样本。根据计算信息增益的定义，梯度大意味着该样本的信息增益大。与 XGBoost 遍历所有特征值的做法相比，避免了不必要的空间上和时间上的开销，进一步提升了模型的高效率。

## 2.4 多变量数据处理方法

在处理具有大量特征或高维度的数据集时，有两种主要策略可以减少数据集的维度。这对于更有效地训练机器学习模型并避免由于特征过多而导致的过拟合至关重要。第一种方法涉及降维技术，例如主成分分析和线性判别分析。这些方法将数据从原始空间投影到新空间，从而降低数据的维度。第二种策略是特征选择，它确定最适合机器学习模型的特征子集。这很重要，因为过多的特征可能会引入噪声、冗余或不相关的信息，这可能会误导模型并阻碍其学习最具预测性的关系。

### 2.4.1 特征降维

在机器学习的背景下，特征降维涉及将数据从原始的高维特征空间转换到一个较低维度的空间，同时尽量保留原有特征空间中的信息。这类方法主要被应用于降低数据处理的复杂度和存储成本。这种技术对于简化模型、提高性能以及加快训练过程至关重要，特别是在处理高维数据时。

#### 2.4.1.1 主成分分析

主成分分析（Principal Component Analysis，PCA）是一种强大的统计技术, 通常用于高维度数据的降维。Pearson[25]提出的论文，被认为是 PCA 的起源，PCA 把原始数据通过线性变换转换成一组各个维度线性无关的形式，可降低原始数据存在的噪音与数据冗余。PCA 背后的核心思想是在尽可能保留数据集中方差最大的方向的前提下，减少由许多相互关联的变量组成的数据集的维度。这通过将原始变量转换为一组新变量来实现，这些新变量称为主成分（principal components）。

第一个主成分具有尽可能高的方差（即尽可能解释数据中的变异性），而每个随后的成分依次具有尽可能高的方差，但受到与前面成分正交的约束。

#### 2.4.1.2 稀疏主成分分析

稀疏主成分分析（Sparse PCA）[26]是传统主成分分析（PCA）的一种变体，它在主成分中引入了稀疏性，使得结果更容易解释，因为每个主成分只涉及一部分特征。稀疏 PCA 在 PCA 解决的优化问题中引入了正则化项（通常是 L1 惩罚，类似于 Lasso 回归），这种惩罚鼓励解决方案是稀疏的，意味着它具有许多零系数。这在数据集具有大量特征，其中许多特征可能对理解数据中的主要变化成分贡献较小的情况下特别有益。

在具有非常多特征的数据集中，传统的 PCA 可能变得难以解释，因为每个主成分都是所有原始特征的线性组合。通过生成仅由少数原始特征的线性组合构成的主成分，稀疏 PCA 有助于在高维环境中保持可解释性。

### 2.4.2 特征选择

特征选择是机器学习中至关重要的过程，涉及挑选一组相关特征（变量、预测因子），用于构建模型。其主要目标是通过消除不必要、无关或冗余的数据来提升模型性能，这些数据可能会使模型效率低、解释性差，并潜在地降低准确性。特征选择的好处体现于随着特征维度的下降，不仅可以降低模型过拟合的风险，模型的训练时间也随之下降。

#### 2.4.2.1 过滤方法

过滤方法使用统计量为每个特征分配一个分数。特征根据分数进行排名，然后选择保留或从数据集中移除。过滤方法通常是单变量的，并且在不依赖模型的情况下独立考虑每个特征。常见的例子包括使用相关系数、卡方检验和互信息得分。本文将对方差阈值过滤法和基于相关性过滤法进行介绍：

1. 方差阈值是一种简单但有效的方法，它移除所有方差未达到某个阈值的特征。因为其理由是方差较低的特征不太可能具有信息性，这种方法旨在消除主要是常数的特征，因此不太可能对模型的预测能力有重大贡献。

2. 基于相关性的单变量特征选择是在构建机器学习模型的预处理阶段筛选特征的一种直接且常用的方法。这种方法涉及使用统计相关性度量来评估每个独立变量（特征）与因变量（目标）之间关系的强度。

#### 2.4.2.2 包裹法

这些方法将选择一组特征视为一个搜索问题，其中准备、评估并比较不同的组合。使用预测模型对每个特征组合进行评分，以确定哪一个在预测目标变量方面最有效。示例包括递归特征消除、前向特征选择和后向特征消除。

首先介绍顺序特征选择器（Sequential Feature Selector, SFS）特征选择技术，它的操作方式是从无特征开始逐个添加（前向选择），或从所有特征开始逐个移除（后向消除），这取决于指定的性能标准。这种方法特别有助于提升模型性能、减少过拟合，并通过消除无关或冗余的特征提高模型的可解释性。

顺序特征选择器的缺点是，特征选择的顺序影响顺序选择器的结果，这是后向和前向顺序特征选择的公认特点。SFS 的本质在于其逐步方法，该方法一次评估一个特征，基于每一步的感知重要性添加或移除特征。这通常会导致路径依赖的结果。还有一种可能性是，SFS 可能会满足于局部最优，而不是全局最优。

其次，递归特征消除（Recursive Feature Elimination, RFE）也是是包裹特征选择方法之一。在需要减少数据维度、增强模型解释性或提高预测建模过程效率的场景中，RFE 尤为有用。

RFE 方法的初始步骤是在数据集中可用的全部特征集上训练模型。模型训练完成后，每个特征都会被赋予一个重要性得分，这通常由模型的系数或特征重要性属性决定。例如，在像线性回归这样的模型中，系数可以指示重要性；而在基于树的模型如决策树或随机森林中，则使用特征重要性。接着，递归地从当前特征集中消除最不重要的特征（得分最低的特征。这一过程重复进行，每一步都使用减少后的特征集重新训练模型，直到达到预定义的停止标准。Guyon 等人[27]展示了在基因选择中使用递归特征消除（RFE）的方法，特别是基于 DNA 微阵列实验中的基因表达数据来选择最能指示癌症的基因。以下是递归特征消除法与交叉验证结合算法的伪代码：

**Algorithm 1** 递归特征消除与交叉验证的包裹法 (RFECV)

**Require:** $X$, $y$: 数据集, $k$: 折叠数量, $s$: 步长, $c$: 成本函数
**Ensure:** 最优特征集
1: 初始化特征集为全部特征
2: **while** 特征数量 > 1 **do**
3: 在当前特征集上训练交叉验证
4: 计算交叉验证成本
5: 移除表现最差的特征
6: **end while**
7: **return** 最优特征集

#### 2.4.2.3 嵌入方法

嵌入式方法涉及在模型训练过程中固有地执行特征选择的算法。这些方法包括自动考虑特征之间的相互作用及其与预测结果的相关性的机制。它们能够这样做是因为它们具有内置的惩罚机制或使用特定标准在模型构建阶段添加或忽略特征。与包裹方法相比，这些方法可以更加计算高效，因为它们不需要为每个特征子集从头开始重新训练模型；选择过程被整合到模型训练中。以下介绍两种较为常见的嵌入方法：

1. Lasso 回归（最小绝对收缩和选择算子）通过其成本函数中的正则化惩罚部分（L1 范数）起作用，这种惩罚具有迫使模型的一些系数精确地变为零的效果。这意味着与这些系数相对应的任何特征实际上都被从模型中排除了。

2. 在决策树模型中，节点的分裂是基于训练数据中能最有效地分离类别或减少回归中的方差的各种特征进行的。不增加预测准确性的特征不会被用于分裂，因此可以有效地被视为已移除。

# 第三章　数据预处理与脑龄预测模型构建

## 3.1　数据预处理

本文所选用的数据集是由全脑结构磁共振图像经过预处理之后得到的表格数据（tabular data）。首先。该数据集的特征数量非常庞大，特征维度为 410（包目标变量）。其次，所使用数据集特征包括性别、年龄、体积、表面积、厚度、平均曲率、高斯曲率以及 MRI 扫描仪类型。其中只有性别与 MRI 扫描仪类型为离散数据（categorical data），其余特征都是连续数据（numerical data）。

数据集由 13 个 comma separated values（csv）文件组成，其中包括左右脑的中各个大脑相关指标如：体积、表面积、高斯曲率、平均曲率等等。每个文件的特征对应着各个指标在大脑不同区域的测量。本文把数据分成三组进行介绍，每组数据特征对应的大脑区域都稍有差异，得到的三组分别为 aseg 数据集，wmparc 数据集与 others 数据集。以下是它们应的业务含义以及说明如下：

**表 3-1　aseg 文件特征含义说明**

| 大脑区域 | 特征 | 功能关联 |
| --- | --- | --- |
| 小脑 | Left-Cerebellum-White-Matter, Left-Cerebellum-Cortex, Right-Cerebellum-White-Matter, Right-Cerebellum-Cortex | 协调姿势、平衡、协调和语言等资源运动。 |
| 脑干 | Brain-Stem | 负责心率、呼吸、睡眠和进食等基本大脑功能。 |
| 基底神经节 | Left-Caudate, Left-Putamen, Left-Pallidum, Right-Caudate, Right-Putamen, Right-Pallidum, Left-Accumbens-area, Right-Accumbens-area | 主要负责运动控制和包括学习在内的认知功能。 |

续表 3-1

| 大脑区域 | 特征 | 功能关联 |
| --- | --- | --- |
| 边缘系统 | Left-Thalamus, Right-Thalamus | 支持情感、行为、动机、长期记忆和嗅觉等多种功能。 |
| 间脑 | Left-Hippocampus, Left-Amygdala, Right-Hippocampus, Right-Amygdala | 包含丘脑和下丘脑等结构，对传递感觉和运动信号至大脑皮层、调节意识、睡眠和警觉状态至关重要。 |
| 视交叉 | Optic-Chiasm | 对双目视觉至关重要。 |
| 胼胝体 | CC_Posterior, CC_Mid_Posterior, CC_Central, CC_Mid_Anterior, CC_Anterior | 在大脑两半球之间整合运动、感觉和认知性能。 |

**表 3-2　wmparc 文件数据特征含义说明**

| 大脑区域 | 特征 | 功能关联 |
| --- | --- | --- |
| 额叶 | bankssts, caudalanteriorcingulate, caudalmiddlefrontal, lateralorbitofrontal, medialorbitofrontal, parsopercularis, parsorbitalis, parstriangularis, precentral, rostralanteriorcingulate, rostralmiddlefrontal, superiorfrontal, frontalpole | 执行功能、运动控制、语言、情绪调节 |
| 顶叶 | inferiorparietal, postcentral, precuneus, superiorparietal, supramarginal | 感觉信息处理、感觉与视觉整合 |
| 颞叶 | bankssts, fusiform, inferiortemporal, middletemporal, parahippocampal, superior temporal, temporalpole, transversetemporal | 听力处理、记忆、语言 |
| 枕叶 | cuneus, lingual, lateraloccipital, pericalcarine | 视觉处理 |

续表 3-2

| 大脑区域 | 特征 | 功能关联 |
| --- | --- | --- |
| 边缘 | isthmuscingulate, parahippocampal | 情感形成与处理、学习、记忆 |
| 岛叶和扣带 | insula, anteriorcingulate, midcingulate | 情绪调节、感知、运动控制、认知功能 |

**表 3-3　others 文件特征含义说明**

| 大脑区域 | 特征 | 功能关联 |
| --- | --- | --- |
| 额叶 | medialorbitofrontal, lateralorbitofrontal, superiorfrontal, rostralmiddlefrontal, caudalmiddlefrontal, precentral, frontalpole, parsopercularis, parstriangularis, rostralanteriorcingulate, caudalanteriorcingulate | 执行功能，决策制定，运动控制，情绪调节 |
| 顶叶 | postcentral, superiorparietal, inferiorparietal, supramarginal | 感觉知觉和整合，空间推理 |
| 颞叶 | superiortemporal, middletemporal, inferiortemporal, temporalpole, transversetemporal | 听力，记忆，言语，情感反应 |
| 枕叶 | cuneus, pericalcarine, lingual, lateraloccipital | 视觉处理 |
| 边缘叶 | parahippocampal, entorhinal, isthmuscingulate, posteriorcingulate, precuneus | 情感，行为，动机，长期记忆 |
| 岛叶 | insula | 情感反应，体内平衡，知觉 |

本文进行模型训练的时候把所有对应文件的数据合并到 original.csv 文件，方便使用 pandas 包进行数据预处理。

### 3.1.1 训练集与测试集划分

在机器学习中，当模型创建时使用了来自训练数据集以外的信息，就会发生数据泄露。这种情况可以发生在对整个数据集进行变换后再执行训练-测试分割的时候。这种泄露会导致过于乐观的性能估计，以及一个在测试数据上表现良好但在真实世界数据上表现不佳的模型。例如：如果你使用整个数据集的分布来进行缩放或标准化，那么模型就会根据包括测试集在内的整体数据分布来调整。这意味着当你训练模型时，它已经根据测试集的特定特征（如均值和方差）进行了调整。所以，在对数据集进行任何数据预处理操作前，一般都得先进性划分，保证最终预测模型在未知数据上的评估是公平的。

本文以 8：2 比例划分训练集与测试集，训练集包括 1280 条数据，测试集 320 条数据，进行模型训练与评估，完成脑龄预测任务。

### 3.1.2 数据标准化

数据标准化是数据预处理阶段的重要步骤之一。在多特征数据集中，当各特征的单位不同（如温度用摄氏度，距离用千米）时，进行标准化能保证这些特征在模型训练过程中具有可比性，避免因单位不同而导致的误解。此外，如果所有特征都按同一尺度标准化，优化算法如梯度下降也可以更有效更快地收敛。这是因为标准化后的特征具有相似的范围，从而使得优化过程更加平滑，减少了某些参数在更新过程中的摆动幅度。

### 3.1.3 类型特征的特征编码

本小节将介绍本文使用的两个类型特征的编码方法：

首先是独热编码（one-hot encoding），对于每个类型特征，独热编码先确定类型特征的唯一类别个数，并为之创建一个新的二进制列。如果数据点中的原始特征属于该类别，那么该数据点对应的列设为 1。所有其他列设为 0。

其次，本文使用了标签编码对类型特征进行编码。标签编码能够简单的将一个类型特征的所有唯一类别映射到自然数，二者之间建立了一一对应关系。该编码也在该类型特征引入了序数性。例如：[poor , mid, excellent] 原本存在序数性的类型特征采用标签编码就可以充分利用标签的特性。

### 3.1.4 特征工程

特征工程是构建机器学习模型过程中的关键步骤。它涉及创建新特征或修改现有特征以提高模型的性能和准确性。

#### 3.1.4.1 基于分箱的特征创建

分箱也被称为离散化，用于减少数据集的噪声，并通过将一组数值分组到箱中，将数值特征转换为类型特征。有几种策略可以对数值数据集执行分箱操作：

1. 等宽分箱：数据的范围被划分为 N 个大小相等的区间，每个区间的宽度为（最大值-最小值）/N。这种方法简单，但如果数据偏斜，可能会导致箱内数据分布不均。

2. 等频分箱：数据被划分为 N 个组，每个组包含大致相同数量的值，因此能更好地处理偏斜数据。这种方法使用分位数来确保数据在箱中的分布更加平衡。

3. 自定义策略分箱：结合领域知识进行分箱。

本文采用分箱操作把所有连续特征转换为类型特征，利用标签编码进行对类型特征编码。

#### 3.1.4.2 分组聚合特征创建

分组聚合特征是一种将数据在预定义组内进行聚合或汇总的方法，旨在揭示可能被单个数据点所掩盖的有意义的见解或模式。这种方法在数据变量之间的关系可能因数据不同部分或组而变化的情景中尤为重要。

本文采用分箱技术，把年龄特征进行离散化，得到年龄的‘分组’。之后根据年龄组，计算各个组内的统计度量如平均值、中位数、标准差等，以了解组内数据的趋势和变异性。再把这些统计度量视为新增特征，融合到原始特征集合。

#### 3.1.4.3 比率和聚合特征创建

本文根据已有的特征创建了大脑聚合特征和比率特征，新增特征如下表所示：

**表 3-4　大脑聚合特征和比率特征**

| 新建特征 | 特征含义说明 |
| --- | --- |
| total_white_vol | 代表大脑中白质的总体积。白质主要由神经纤维组成，这些纤维有助于不同大脑区域之间的通信。 |
| total_gray_vol | 指大脑中灰质的总体积，主要包括神经元细胞体，对于大脑中的信息处理至关重要。 |
| total_brain_vol | 灰质、白质及其他大脑结构的总体积的综合。这是衡量大脑整体大小的指标。 |
| total_surface_area | 大脑皮层的总表面积。这一指标反映了皮层的折叠性质，表面积较大通常表明皮层折叠更多。 |
| total_thickness | 大脑整体的平均皮层厚度。这个指标可以指示大脑组织的紧凑程度，并且已与多种认知和健康状况关联。 |
| gray_to_white_ratio | 灰质体积与白质体积的比率。这个比率有助于评估大脑的组成以及可能随疾病或老化发生的变化。 |
| white_to_gray_ratio | 白质体积与灰质体积的比率。这是大脑组成的另一种表示方式，提供了从另一个角度的见解。 |
| left_to_right_ratio | 比较大脑左半球与右半球的体积或大小。这个比率在研究大脑功能的侧化方面非常重要。 |
| volumes_surface_ratio | 总大脑体积与总表面积的比率。它可以提供关于大脑的物理属性的见解，如密度和折叠模式。 |
| curvature_ratio | 测量大脑表面的弯曲程度。这一特征有助于识别大脑区域内典型或非典型的折叠模式。 |
| volume_gauss_ratio | 这可能指的是涉及体积测量和从大脑表面的高斯曲率得出的测量的比率，反映复杂的几何属性。 |
| thickness_volume_ratio | 皮层厚度与总大脑体积的比率，提供了厚度相对于大脑大小的变化的视角。 |

#### 3.1.4.4 特征选择

本文所采用递归特征消除与交叉验证（RFECV），进行自动的特征筛选。RFECV 通过将交叉验证纳入基本的递归特征消除（RFE）方法中，有效地优化模型的特征数量，以提高性能并避免过拟合。当数据维度太大时，本文将采用嵌入方法配合 LightGBM 模型进行特征选择。除此之外，本文也使用了方差阈值和基于相关性的单变量特征选择降低特征维度。

## 3.2 预测模型构建

首先，通过对数据进行预处理，包括标准化和特征选择，优化了输入数据的质量。接着，构建预测模型，完成大脑年龄预测任务。

### 3.2.1 线性回归模型构建

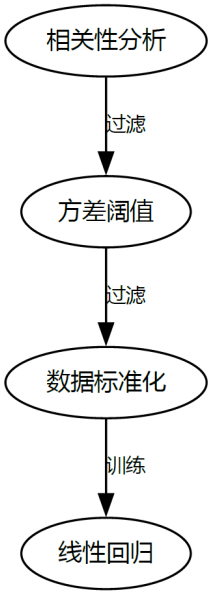


**图 3-1 线性回归模型 pipeline**

首先，本文对已经过数据预处理的数据集用线性回归模型进行预测。线性回归通常被认为是统计学和机器学习中回归任务最基础和基本的模型。因其简单性和可解释性，线性回归成为理解回归分析的 baseline 模型。

在本研究中，为了有效降低数据的复杂度并提高模型的解释性与运算效率，我们将采用一系列特征选择方法，如方差阈值法（Variance Thresholding）和相关

性分析。这些方法能够在模型训练前去除冗余特征，从而降低数据的复杂度。

ElasticNet 模型结合了 L1 和 L2 正则化，有效于处理高维数据中的共线性问题和进行特征选择。然而，由于本文已通过预处理步骤实施了方差阈值法和相关性分析，这些方法本身就旨在减少特征间的多重共线性并选出最有影响力的特征。因此，再使用 ElasticNet 进行正则化可能会导致过度约束，使模型的灵活性受限，且可能不会带来额外的性能提升。

本文使用 sklearn 机器学习包定义了一个 pipeline，pipeline 包含了方差阈值法（Variance Thresholding）和相关性分析图 3-1 给出。

### 3.2.2 创建 GBDT 模型

本文不仅构建了基于线性回归的预测模型，还选用了当前性能极为强大的 XGBoost 和 LightGBM 模型进行脑龄预测。线性回归模型虽然在处理线性关系时表现出色，但它在捕捉数据特征之间的非线性关系方面存在局限性。相比之下，XGBoost 和 LightGBM 作为一种梯度提升框架，通过构建决策树来有效地揭示特征间的非线性关系和复杂的相互作用。

线性回归模型在许多情况下提供了快速和解释性强的解决方案，尤其是在特征和响应变量之间关系较为直接和简单时。然而，本文数据集由医学影像数据经过预处理得到的表格数据更为复杂，单纯的线性模型往往难以表现出最佳性能。在这种情况下本文采用基于 GBDT 的模型希望可以提供更精确的预测结果，也能通过其内置的特征重要性评估帮助我们理解哪些变量对预测结果影响最大。因此，本研究采用了双模型策略，即通过线性回归模型建立基线性能，并通过 GBDT 模型探索和利用数据的非线性特征。

本文分别使用 xgboost 和 lightgbm 机器学习包里的 XGBRegressor 和 LGBM-Regressor 函数，构建了预测脑龄的机器学习模型。

### 3.2.3 贝叶斯调参

超参数调参是机器学习模型开发中的一个重要环节，目的是找到最佳的超参数设置，以最大化模型的性能。超参数调参一般有几种策略可以采用：

首先，网格搜索（Grid Search）是一个穷举搜索方法，先定义一个参数搜索

空间，然后网格搜索会对每一组参数进行训练，之后用交叉验证评估每组参数地效果。该方法简单易懂，但其开销非常庞大，适用于参数空间较小的预测模型。

其次，随机搜索（Random Search）从已定义的参数搜索空间随机选取参数组进行训练与评估，显然该方法与网格搜索相比，更为高效。

贝叶斯调参与前者相比是一种更为高级的方法，使用了利用贝叶斯统计原理来选择最佳的超参数组合，从而优化了预测模型的性能。其核心方法是使用高斯过程模型建立了超参数与目标函数之间的关系。高斯过程提供了对未知目标函数的平滑估计，并能够在每个点上给出平均值和不确定性的估计。高斯过程作为目标函数的代理模型（surrogate model）：

$$f(\theta) \sim \mathcal{GP}(m(\theta), k(\theta, \theta')), \tag{3-1}$$

其中 $m(\theta)$ 是均值函数，$k(\theta, \theta')$ 是协方差函数，描述超参数空间中不同点之间的相关性。

当新的数据点（即超参数组合及其对应的性能评估）可用时，高斯过程的先验分布会根据这些数据点更新，成为后验分布。后验分布反映了在观测到新数据后，对目标函数的改进估计和不确定性的减少。之后是采集函数的定义，常见的采集函数包括预期改进量（Expected Improvement）、概率改进（Probability of Improvement）和知识梯度（Knowledge Gradient）。一旦定义了采集函数，下一步是在超参数空间中找到最大化采集函数的参数组合。采集函数 $a(\theta)$ 指导选择下一组要评估的超参数：

$$\theta_{\text{next}} = \arg\max_{\theta} a(\theta|\mathcal{D}), \tag{3-2}$$

其中 $\mathcal{D}$ 表示当前的观测集。

选定的超参数组合预计会带来最大的性能提升，或者最大的信息增益，帮助更好地理解目标函数的行为。使用选定的超参数组合来训练模型并评估其性能，然后将这些新的评估结果反馈到高斯过程中，更新后验分布。这个过程不断重复，每次迭代都在细化我们对目标函数的理解，并逐步逼近最优的超参数组合。

#### 3.2.3.1 XGBoost 超参数

在使用 XGBoost 进行小型数据集建模时，适当调整超参数至关重要，以确保模型不会因过度拟合训练数据而失去泛化能力。以下是一些关键超参数的详细解释，以及如何调整它们来防止过拟合。

max_depth 参数决定了树的深度。较深的树能够学习到更复杂的数据模式，但也更容易过拟合。在小数据集中，推荐将 max_depth 设置在较低的水平（如 3 到 10）之间，这有助于简化模型结构，同时避免学习到数据中的噪声。

min_child_weight 参数控制在决策树的节点进行分裂所需的最小权重和。设置较高的 min_child_weight 可以避免模型在数据中的小波动上做出过度反应，从而防止过拟合。这样，只有当找到一个具有统计意义的分裂点时，模型才会进行分裂。

gamma 参数指定了节点分裂所需的最小损失减少量。增加此参数的值会使模型的生长更为保守，减少树的复杂度。这有助于控制模型的过度拟合，因为它通过阻止那些不会带来显著改善的分裂来增强模型的泛化能力。

subsample 参数定义了用于构建每棵树的数据的抽样比例。较低的比例可以增加训练过程中的随机性，防止过拟合。通常建议将 subsample 设置在 0.5 到 0.8 之间，以此来平衡模型的偏差和方差。

colsample_bytree 参数控制每棵树的特征采样比例。通过调整这些参数小于 1，可以确保每次分裂不会使用所有的特征，从而增加模型的泛化能力，并减少过拟合风险。

lambda 是 L2 正则化项，增加此值可以减轻模型的过拟合问题，因为它惩罚模型的复杂度；而 alpha 是 L1 正则化项，它有助于使模型更加稀疏，可以将不重要特征的系数减至零。适当增加这些正则化参数有助于提升模型的泛化能力。

n_estimators 参数代表在 XGBoost 模型中构建的树的数量。在决定最终模型复杂性和能力方面，这是一个关键参数。增加树的数量可以帮助模型更好地学习和拟合数据，但同时也增加了过拟合的风险。对于小型数据集，过多的树可能会导致模型过度学习数据中的噪声和细节。

#### 3.2.3.2 LightGBM 超参数

在使用 LightGBM 模型处理小型数据集时，合理调整超参数是至关重要的，因为这能有效防止模型过拟合并提高预测的泛化能力。本小节将介绍了若干关键超参数，并提供了针对每个参数的调整建议，以期在保持模型性能的同时，减少过拟合的风险。

max_depth 参数控制每棵树的最大深度。在决策树中，深度较大的树能够模拟更加复杂的数据模式，但同时也增加了过拟合的风险。通过合理调整此参数，可以有效平衡模型的偏差与方差，使模型既能学习到数据的关键特征，又不会过度学习训练数据中的噪声。

feature_fraction 参数用于设置在每次迭代中随机选择一定比例的特征进行训练，这是一种特征子采样的方法。此举不仅可以减少计算资源的消耗，加快模型的训练速度，还能通过增加训练过程的随机性来提高模型的泛化能力，降低过拟合的可能性。

n_estimators 定义了构建的树的数量，即模型的提升轮数。树的数量越多，模型的复杂度越高，理论上可以捕获更多的数据细节。但是，树过多也可能导致模型学习过度，从而对训练数据过拟合。因此，选择一个适中的树数量对于确保模型既能达到较高的准确率又具有良好泛化能力至关重要。

min_data_in_leaf 是决定一个叶子节点所需要的最小样本数，这个参数在回归任务中尤其重要。设置较大的 min_data_in_leaf 值可以防止模型在训练数据上创建过于复杂的树结构，从而避免模型在特定数据点上过度拟合。这有助于增加模型的健壮性，减少由于训练数据中随机噪声导致的误差。

num_leaves 决定了每棵树中叶子的最大数量，与 max_depth 类似，叶子节点越多，模型可以学习到更多的数据细节，但同时也增加了模型的复杂性。如果未与 min_data_in_leaf 等参数相结合，过多的叶子可能会导致模型结构过于复杂，从而引起过拟合问题。

wlambda_l2 是 L2 正则化项，lambda_l1 是 L1 正则化项，分别对应于 XGBoost 的 lambda 和 alpha。

## 3.3 模型性能指标

大脑年龄预测的模型一般都是回归模型，实施能够准确反映模型预测性能的健壮性指标是至关重要的。本文将使用平均绝对误差（Mean Absolute Error，MAE）、皮尔逊相关系数（Pearson Correlation Coefficient，PCC）和决定系数（$R^2$）作为模型性能的评估指标。

平均绝对误差是一个直接的度量标准，用于衡量一组预测中误差的平均大小，而不考虑其方向。它是预测值与观测值之间绝对差的平均值。在大脑年龄预测的背景下，MAE 提供了一个清晰直观的度量，表明模型根据 MRI 数据或其他相关生物标记预测大脑的年龄的准确性。较低的 MAE 值表明模型能更准确地预测实际大脑年龄，这对于神经疾病的早期检测和干预至关重要。平均绝对误差（MAE）计算如下：

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{3-3}$$

其中 $n$ 是样本数量，$y_i$ 是第 $i$ 个样本的实际值，$\hat{y}_i$ 是第 $i$ 个样本的预测值。

皮尔逊相关系数对预测的大脑年龄与真实年龄之间的线性相关性进行评估。这个系数的范围从-1 到 1，其中 1 表示完美的正线性关系，-1 表示完美的负线性关系，接近零的值表示没有线性相关。在大脑年龄预测模型中，高的皮尔逊相关值表明模型预测与实际年龄一致地增加。设有两个变量 $X$ 和 $Y$，各自拥有 $n$ 个观测值。皮尔逊相关系数 $r$ 可通过以下公式计算：

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \tag{3-4}$$

其中，$\overline{X}$ 和 $\overline{Y}$ 分别是 $X$ 和 $Y$ 的样本均值。

$R^2$ 指标，常被称为决定系数，代表了模型中因变量可由自变量解释的方差百分比。在大脑年龄预测中，更高的 $R^2$ 值表明模型可以解释实际大脑年龄的方差的重要部分，突出了其有效性。$R^2$ 在比较不同模型的性能或评估同一模型经调整后的改进方面特别有用，因为它提供了一个衡量模型关于大脑年龄变化解释力的标尺。决定系数的计算公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3-5}$$

其中：$y_i$ 是第 $i$ 个观测值，$\hat{y}_i$ 是第 $i$ 个观测值的预测值，$\bar{y}$ 是所有观测值的平均值，$n$ 是观测值的总数。$R^2$ 的值通常介于 0 到 1 之间：当 $R^2 = 1$ 时，表明模型完美地预测了所有观测值。当 $R^2 = 0$ 时，表明模型的预测与观测值的平均值一样好。当 $R^2$ 为负值时，表明模型的预测比使用观测值的平均值还要差。通常，$R^2$ 越接近 1，表明模型的预测能力越强，拟合效果越好。

这些指标共同提供了一个全面的大脑年龄预测模型评估框架。MAE 直接衡量误差大小，PCC 提供关于预测年龄与实际年龄一致性的洞察，$R^2$ 评估模型的整体解释力。结合使用这些指标可以对模型的能力进行健全的评估，指导改进并验证其在临床和研究应用中的实用性。这种方法确保模型不仅准确预测大脑年龄，而且以符合生物学预期的方式进行，增强了其在诊断和监测与年龄相关的神经疾病条件的实际价值。在本文中，我们选择以平均绝对误差为主要评估指标，同时辅以决定系数（$R^2$）和皮尔逊相关系数作为辅助指标，以全面评估模型的预测能力和统计特性。

# 第四章　实验结果和分析

## 4.1　实验流程

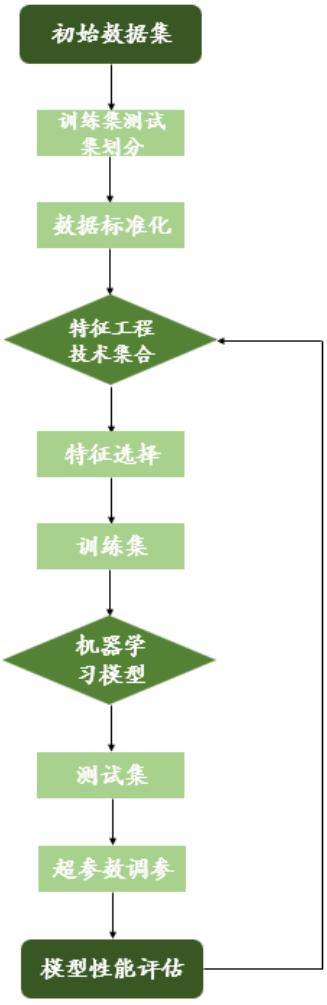


**图 4-1　实验流程**

图 4-1 展示了本文中研究的模型开发流程，特别强调了特征工程在模型构建中的重要性。整个流程从“数据准备”阶段开始，涉及数据的收集和初步处理。首先，数据被划分为训练集与测试集，并对它们进行标准化处理，以确保输入数据的一致性和可比性。

其次，流程进入“数据特征化”阶段。在这一阶段，文章主要通过比较不同

的特征工程方法来探索它们如何影响模型的最终性能。这一步骤包括了广泛的特征提取和选择过程，是筛选出最有效模型的关键环节。

然后，所选特征用于“训练模型”。本文采用多种算法来训练模型，并比较这些模型在不同特征工程技术下的表现。模型训练完成后，进入评估阶段，在此阶段，模型通过一系列指标进行性能评估，目的是确保找到在特定特征工程方法下表现最佳的模型。本文不仅为模型开发提供了清晰的视图，而且突出了进行恰当特征工程的重要性，以确保模型达到最优性能。

## 4.2 实验结果分析

本文实验阶段首先以线性回归模型模型为 baseline 模型进行脑龄预测任务。baseline 模型在各个模型性能指标下如下表：

**表 4-1 baseline 模型性能**

| 模型 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 线性回归 | 10.54907 | 9.46210 | 0.03638 | 0.41650 |

本文将探索各种特征工程技术，并评估这些技术对基准模型在预测任务中性能的影响。研究的主要目的是通过实施有效的特征工程方法来优化模型性能。本文尝试诸多特征工程技术，但大多数都对模型性能的提高都微乎其微。本章只讨论成功优化预测模型性能的特征工程方法。

### 4.2.1 单变量过滤法对模型预测性能结果分析

从基本的线性回归模型到包含方差阈值和相关性分析的模型，MAE、$R^2$ 和皮尔逊相关系数的逐步改进表明，特征选择在提升模型性能中起着关键作用。通过移除较少信息或冗余的特征，模型似乎能更好地捕捉数据中的底层模式，从而提高预测准确性。

**表 4-2 单变量过滤法线性回归结果**

| 模型 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 线性回归 | 10.54907 | 9.46210 | 0.03638 | 0.41650 |
| 线性回归 + 方差阈值 | 10.47238 | 9.33832 | 0.06327 | 0.43001 |

续表 4-2

| 模型 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 线性回归 + 相关性分析 | 9.84975 | 9.08579 | 0.17030 | 0.47343 |

以上结果显示，所有模型的 $R^2$ 值偏低，这表明经过了这些相对简单的特征筛选，其解释因变量方差的能力依然有限，可从以上结果得到几个结论：

1. 线性回归模型可能过于简化，不足以描述数据集中存在的复杂关系。数据可能含有非线性模式，而这些模式无法通过简单的线性方法有效捕捉。可考虑更为复杂的非线性回归模型，如 LightGBM，可助于提高预测精度和可解释性。

2. 目前模型所使用的特征集可能未能全面捕捉到影响因变量的所有重要因素。考虑使用更为高级的特征工程技术如主成分分析进行实验。

### 4.2.2 主成分分析特征降维方法结果分析

**表 4-3 主成分分析线性回归模型结果**

| 比例 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 0.80 | 9.74463 | 8.74186 | 0.24079 | 0.49448 |
| 0.85 | 9.67844 | 8.71840 | 0.24313 | 0.49906 |
| **0.90** | **9.55827** | **8.72534** | **0.20942** | **0.49210** |
| 0.95 | 9.57127 | 8.83460 | 0.11847 | 0.45763 |

注：当 n_components 是一个浮点数（例如 0.9）时，PCA 将选择所需的最小成分数量，以解释至少该比例（本例中为 90%）的数据总方差。

从以上表中，我们可以观察到一些关键点：

1. 当配置 n_components = 0.90 实现了最低的交叉验证 MAE，表明它在多个验证集中最小化了平均预测误差。然而，其 $R^2$ 得分低于 0.80 和 0.85 解释方差的配置，这表明在因变量的方差解释方面拟合较差。

2. 当 n_components = 0.95 导致 MAE 和 $R^2$ 的性能都变差，可能是因为增加 n_components 捕获了更多噪声而非有用的方差。

**表 4-4　主成分分析 XGBoost 模型结果**

| 比例 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 0.80 | 10.58231 | 9.54913 | 0.08654 | 0.35968 |
| 0.85 | 10.64801 | 9.81418 | 0.08756 | 0.36340 |
| 0.90 | 10.70046 | 9.60013 | 0.09017 | 0.36848 |
| 0.95 | 10.57 | 9.79707 | 0.07440 | 0.33299 |

结果表明，保留更多特征（较高方差比例）并不一致地改善模型性能，各种指标在最高方差比例前达到峰值或显示最优值。这可能表明通过更高方差比例保留的额外特征引入了噪声或无关信息，这使得 XGBoost 模型混淆而不是帮助它。

**表 4-5　主成分分析 LightGBM 模型结果**

| 比例 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 0.80 | 10.10219 | 9.31355 | 0.13686 | 0.39864 |
| 0.85 | 10.08565 | 9.19015 | 0.17660 | 0.43395 |
| **0.90** | **9.98732** | **8.99122** | **0.22352** | **0.47330** |
| 0.95 | 10.03308 | 9.30683 | 0.18095 | 0.42845 |

从以上表中，我们可以观察到一些关键点：

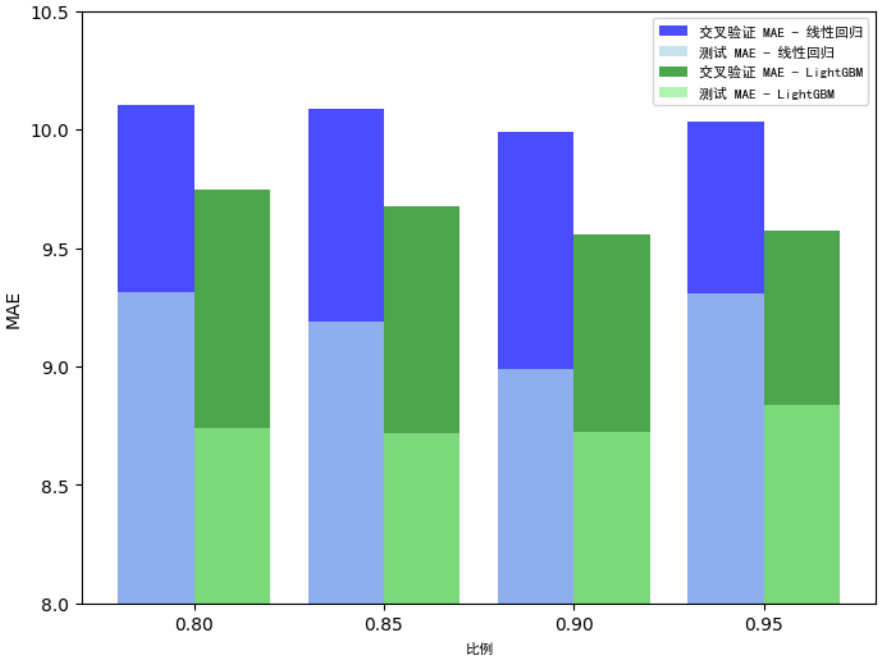


**图 4-2　线性回归模型与 LightGBM 模型在不同 PCA 方差比例下的性能对比**

1. 当组件数 n_components 从 0.80 增加到 0.90 时，所有指标普遍呈现改善趋

势。MAE 降低，表示预测精度提高，$R^2$ 和皮尔逊系数也增加，表明模型拟合效果更好，与实际值的线性关系更强。

2. 当 n_components 增加到 0.95 时，性能明显下降。与 0.90 相比，$R^2$ 和皮尔逊指标都有所下降，MAE 也略有恶化，这可能表明超过 0.90 的额外组件开始引入噪声或较少的信息性方差。

3. n_components = 0.90 的设置显示出最佳的整体性能。

PCA 有效地减少了数据集中的噪声或信息较少的变量。降低维度可能有助于缓解过拟合或计算效率低下等问题，同时仍保留了模型训练所需的关键信息。

### 4.2.3 分组聚合特征创建结果分析

**表 4-6 分组聚合特征 + 线性回归模型结果**

| 年龄分段数 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 4 | 3.87439 | 3.54719 | 0.88608 | 0.94307 |
| 8 | 2.16225 | 1.75991 | 0.96864 | 0.98422 |
| 16 | 1.00290 | 0.90436 | 0.99166 | 0.99583 |
| **32** | **0.54558** | **0.49474** | **0.99733** | **0.99867** |

**表 4-7 分组聚合特征 +XGBoost 模型结果**

| 年龄分段数 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 4 | 3.17879 | 3.33583 | 0.90047 | 0.94951 |
| 8 | 1.78498 | 1.65641 | 0.97366 | 0.98681 |
| 16 | 0.89791 | 0.78408 | 0.99341 | 0.99674 |
| **32** | **0.42192** | **0.39318** | **0.99796** | **0.99897** |

**表 4-8 分组聚合特征 +LightGBM 模型结果**

| 年龄分段数 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 4 | 3.13473 | 3.15839 | 0.90931 | 0.95404 |
| 8 | 1.72951 | 1.64564 | 0.97271 | 0.98747 |
| 16 | 0.85102 | 0.78050 | 0.99324 | 0.99713 |
| **32** | **0.41190** | **0.38153** | **0.99755** | **0.99887** |

以上是线性回归模型、XGBoost 模型和 LightGBM 模型在大脑年龄预测任务中的实验结果。随着年龄分段数增加，各种性能指标都显示出显著的提升。

如果模型在训练数据上表现良好但在未见数据（验证或测试数据）上表现不佳，则表示过拟合。随着分段数的增加而持续改善，并且数值保持接近，没有明显的差异表明过拟合。所有指标（MAE，$R^2$ 和皮尔逊相关系数）随分段数的增加而改善。这种在训练和验证指标上的均匀提升表明模型有效地学习了更多有用的模式，而不是学习了噪声。

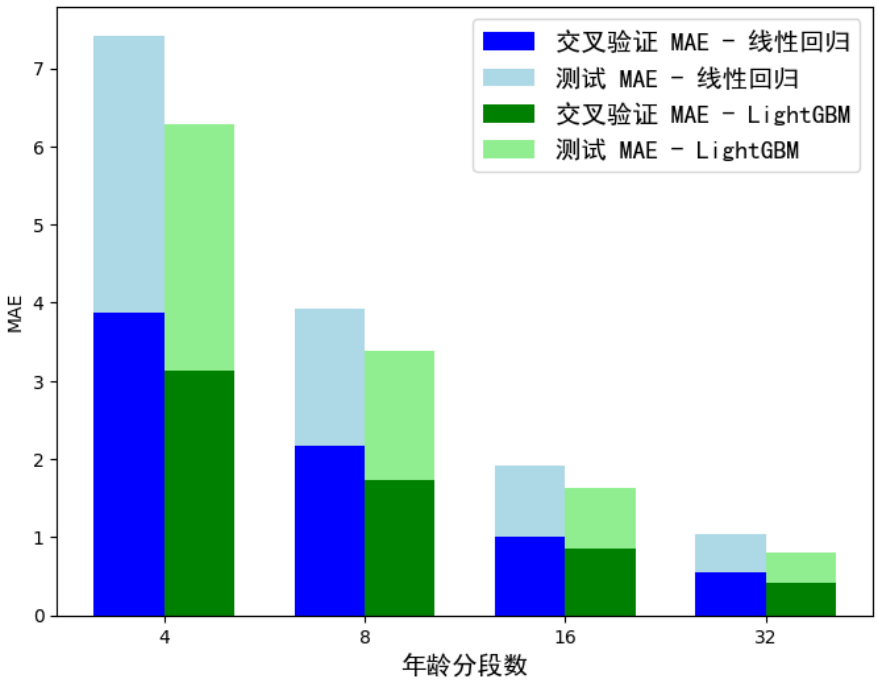


**图 4-3　线性回归模型与 LightGBM 模型在不同年龄分段数下的性能对比**

增加分段数允许更精细的年龄分类，这似乎有助于模型捕捉与大脑老化相关的更详细和重要的模式。这种粒度似乎提供了更精确的针对性，能够更好地捕捉大脑结构或功能的年龄相关变化，这对于准确的大脑年龄预测至关重要。

鉴于模型性能指标随着分段数的增加而改善，并且没有明显的过拟合迹象，使用更细的年龄分段结合分组统计特征的方法对于大脑年龄预测任务来说非常有效。随着模型或数据预处理步骤的不断优化和增加复杂性而可能出现的过拟合问题，必须持续监控模型性能以防模型过拟合。

### 4.2.4　贝叶斯调参结果分析

通过综合考虑这些超参数的设置与调整，可以显著提高模型对新数据的预测能力，同时避免在特定数据集上的过度拟合。这种精细化的参数调优是构建高效、可靠机器学习模型的关键步骤。

**表 4-9　XGBoost 进行贝叶斯调参结果**

| 调参 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 之前 | 0.42192 | 0.39318 | 0.99796 | 0.99897 |
| **之后** | **0.38263** | **0.35772** | **0.99759** | **0.99890** |

以上为经过贝叶斯调参之后的 XGBoost 模型性能指标结果。最终得到的最佳参数为 (colsample_bytree: 1.0), (learning_rate: 0.08752457697622658), (max_depth: 3), (min_child_weight: 6), (n_estimators: 500), (reg_alpha: 0.743804994822914), (reg_lambda: 1.0), (subsample: 1.0)。

**表 4-10　LightGBM 进行贝叶斯调参结果**

| 调参 | 交叉验证 MAE ↓ | 测试 MAE ↓ | $R^2$ ↑ | Pearson ↑ |
|---|---|---|---|---|
| 之前 | 0.41190 | 0.38153 | 0.99755 | 0.99887 |
| **之后** | **0.36824** | **0.35770** | **0.99815** | **0.99903** |

以上为经过贝叶斯调参之后的 LightGBM 模型性能指标结果。最终得到的最佳参数为 (max_depth: -1), (feature_fraction: 0.5), (n_estimators: 256), (min_data_in_leaf: 200), (learning_rate: 0.06073005652705612), (num_leaves: 2), (lambda_l1: 0.1), (wlambda_l: 20.1)。

## 4.3　本章小结

本章详细比较了不同特征工程技术对预测模型性能的影响。本章展示了特征工程在提高模型预测精度中的关键作用，特别是分组聚合特征的使用显著提高了模型的表现。此外，实验结果还显示，LightGBM 模型在所有测试中表现最佳，有效利用了经过特征工程处理的数据，从而达到了更高的预测精度。

综合所有实验结果，我们可以得出以下结论：1) 特征工程是提升模型性能的有效手段，尤其是在处理高维数据时；2) 在所有测试的模型中，基于梯度提升的 LightGBM 模型在特征工程优化后，展示出了最优的性能；3) 实验中使用的评估指标，如 MAE 和 Pearson 相关系数，证实了特征工程的正面影响，尤其是在提升模型对数据的解释能力方面。

# 第五章　结论和展望

## 5.1　工作结论

本研究旨于通过健康成年人 MRI 影像数据（经过预处理得到的表格数据），应用解决高维度数据的特征工程技术，建立一个脑龄预测模型。本文结合不同特征工程技术配合线性回归和梯度提升决策树（LightGBM）模型进行预测实施。

在特征工程阶段，本文尝试了不少方法，包括但不仅限于：分箱技术、基于比率的特征创建、基于分组聚合特征、主成分分析降维等等。其中最为有效的方法为基于分组聚合特征，通过把年龄分成年龄段，在计算各个年龄段的聚合特征，大幅度的提高了预测模型性能。

本文最终采用 LightGBM 为最终模型。LightGBM 无论是在模型效率还是模型效果都非常优秀，在高维度数据上也得到了稳定的发挥。

结合实验过程和实验结果，本文得出以下结论：

1. 在大脑年龄预测模型的开发中，高质量的特征工程发挥着关键作用，这主要得益于其能够识别出原始数据中不直接显现的特征间的潜在关联。这种数据的简化处理有助于使预测模型集中关注更为宏观的趋势，避免模型因响应那些不影响预测准确性的微小波动而产生误导。

2. 随着年龄分段数增加，模型的 MAE 指标不断向零逼近，而且这个收敛速度逐渐放变慢。本文已使用交叉验证法，验证了预测模型随着年龄段的增加也学习到更细粒度信息，从而得到模型性能的提高。然而，本文在进行超参数调参的时候，对树的深度，特征比例等超参数，防止模型的过拟合问题。

3. 本文提出的模型不仅保证了了成年人脑年龄预测的准确性，还实现了数据处理的自动化，显著降低了传统脑年龄预测方法在时间消耗和对专家知识依赖方面的需求。通过精准和快速的分析能力，该模型能够协助医生和科研人员深入理解脑老化过程及其对人类健康的复杂影响，从而推动神经科学和老年医学的发展。

## 5.2 工作展望

本研究主要考察了各种特征工程技术对大脑年龄预测模型性能的提升作用。通过融合神经科学和老年学等领域的专业知识，我们有望进一步增强模型的预测精度。该方法不仅应用统计技术优化了数据的表述，还可以结合专家深度的见解，挖掘出对精确预测大脑年龄至关重要的生物学相关变量及其相互作用。

从模型构建的角度，若之后可以得到更多的数据集，我们可以考虑尝试使用神经网络进行脑龄预测。由于本文所使用的数据集样本数有限，本文决定不考虑使用深度学习进行脑龄预测任务。接着，模型融合也是一个可以进一步提高模型性能的方向。

# 参考文献


[1] RAICHLE M E. Functional brain imaging and human brain function[J]. Journal of Neuroscience, 2003, 23(10): 3959-3962.

[2] FRANKE K, GASER C. Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained?[J]. Frontiers in neurology, 2019, 10: 454252.

[3] SMITH S M, VIDAURRE D, ALFARO-ALMAGRO F, et al. Estimation of brain age delta from brain imaging[J]. Neuroimage, 2019, 200: 528-539.

[4] BAECKER L, GARCIA-DIAS R, VIEIRA S, et al. Machine learning for brain age prediction: Introduction to methods and clinical applications[J]. EBioMedicine, 2021, 72.

[5] RAN C, YANG Y, YE C, et al. Brain age vector: A measure of brain aging with enhanced neurodegenerative disorder specificity[J]. Human brain mapping, 2022, 43(16): 5017-5031.

[6] SIHAG S, MATEOS G, MCMILLAN C, et al. Explainable brain age prediction using covariance neural networks[J]. Advances in Neural Information Processing Systems, 2024, 36.

[7] RAZ N, RODRIGUE K M. Differential aging of the brain: patterns, cognitive correlates and modifiers[J]. Neuroscience & Biobehavioral Reviews, 2006, 30(6): 730-748.

[8] BERNARD B, GOLDMAN J G. MMSE-Mini-Mental State Examination[G]//Encyclopedia of movement disorders. Elsevier Inc, 2010: 187-189.

[9] NASREDDINE Z S, PHILLIPS N A, BÉDIRIAN V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment

[J]. Journal of the American Geriatrics Society, 2005, 53(4): 695-699.

[10] MEMÓRIA C M, YASSUDA M S, NAKANO E Y, et al. Brief screening for mild cognitive impairment: validation of the Brazilian version of the Montreal cognitive assessment[J]. International journal of geriatric psychiatry, 2013, 28(1): 34-40.

[11] LEE J Y, LEE D W, CHO S J, et al. Brief screening for mild cognitive impairment in elderly outpatient clinic: validation of the Korean version of the Montreal Cognitive Assessment[J]. Journal of geriatric psychiatry and neurology, 2008, 21(2): 104-110.

[12] JACK C R, KNOPMAN D S, JAGUST W J, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers[J]. The lancet neurology, 2013, 12(2): 207-216.

[13] BABILONI C, VISSER P J, FRISONI G, et al. Cortical sources of resting EEG rhythms in mild cognitive impairment and subjective memory complaint[J]. Neurobiology of Aging, 2010, 31(10): 1787-1798.

[14] PETERSEN R C. Mild cognitive impairment as a diagnostic entity[J]. Journal of internal medicine, 2004, 256(3): 183-194.

[15] KRÄMER C, STUMME J, da COSTA CAMPOS L, et al. Prediction of cognitive performance differences in older age from multimodal neuroimaging data [J]. GeroScience, 2024, 46(1): 283-308.

[16] COLE J H, POUDEL R P, TSAGKRASOULIS D, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker [J]. NeuroImage, 2017, 163: 115-124.

[17] JÓNSSON B A, BJORNSDOTTIR G, THORGEIRSSON T, et al. Brain age prediction using deep learning uncovers associated sequence variants[J]. Nature communications, 2019, 10(1): 5409.

[18] TANVEER M, GANAIE M, BEHESHTI I, et al. Deep learning for brain age estimation: A systematic review[J]. Information Fusion, 2023.

[19] HOERL A E, KENNARD R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[20] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996, 58(1): 267-288.

[21] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005, 67(2): 301-320.

[22] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. Annals of statistics, 2001: 1189-1232.

[23] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[24] KE G, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.

[25] HOTELLING H. Analysis of a complex of statistical variables into principal components.[J]. Journal of educational psychology, 1933, 24(6): 417.

[26] ZOU H, HASTIE T, TIBSHIRANI R. Sparse principal component analysis[J]. Journal of computational and graphical statistics, 2006, 15(2): 265-286.

[27] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46: 389-422.

# 致　　谢

首先，我要感谢我的毕业论文导师高阳老师，在研究院的学习期间，我从研究院老师、学长学姐那里学到了许多宝贵的知识。同时，特别感谢吴志平学长和 Henry 学长的耐心指导，他们的帮助使我能够顺利完成我的毕业论文。感谢南京大学在过去六年里对我的培养，以及我的父母对我始终不渝的支持和包容。此外，我还要感谢所有在学业上给予我帮助的同学们，感谢你们的陪伴和支持。