**Deciphering Abalone Ages through Machine Learning Methods**

Part 1 Introduction

1.1 Background

传统方法测定鲍鱼的年龄是通过从圆锥体切开鲍鱼壳、染色并通过显微镜计算环数来确定的，这种方法耗时且破坏性，依赖于手工，易受技术人员主观判断影响，且对鲍鱼可能造成伤害，这些因素限制了其效率和准确性，促使研究人员探索更优解决方案。而其他更容易获得的测量结果可用于预测年龄，可能需要更多信息，例如天气模式，位置和​​食物供应情况等。

在本篇报告中，我使用包含鲍鱼的物理测量数据（如长度，高度，总重量等）的UCI Abalone数据集以及R tidyverse构建机器学习模型，以一个更为优雅的方式对鲍鱼年龄进行预测。

1.2 Data Set Introduction

下面的表格是对数据集的一些描述，其中…….是预测变量，目标变量Age并没有出现在数据集中，但是其数值等于Rings+1.5。

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Type | Description | Units |
| Sex | Categorical | M, F, and I (infant) |  |
| Length | Continuous | Longest shell measurement | mm |
| Diameter | Continuous | perpendicular to length | mm |
| Height | Continuous | with meat in shell | mm |
| Whole\_weight | Continuous | whole abalone | grams |
| Shucked\_weight | Continuous | weight of meat | grams |
| Viscera\_weight | Continuous | gut weight (after bleeding) | grams |
| Shell\_weight | Continuous | after being dried | grams |
| Rings | Integer | +1.5 gives the age in years |  |

1.3 Model in Use

报告中使用到了基于Lp正则化的线性回归以及随机森林这两类机器学习算法对处理后的训练集进行训练，最终通过在测试集上的表现比较两个模型的性能。

Pic of Abalone

Part 2 Data Pre-processing and EDA

Part 2，3和4中使用到的数据集组成详见Appendix

2.1 Data Pre-processing

2.1.1 Data Cleaning

由于数据集中除了Sex列其他均为连续型数据，经过检查数据集中不存在Typo错误。同时使用any(is.na(dataset))的时候也没有发现数据中存在缺失值。

执行机器学习任务之前，检查数据集很有必要，训练数据中如果包含有歧义的数据，会导致模型训练出现偏差。于此同时，我发现了一些数据存在逻辑错误，并使用tidyverse中的filter将这些数据集中的这些行删除：

1.数据中存在2行数据中的Height等于0的情况，而鲍鱼的高度不可能等于0；

2.理论上鲍鱼总重量应该等于各部分重量之和，但是有相当一部分行（154 rows）存在WholeWight < ShuckedWeight + ShellWeight的情况。

2.1.2 Add Age Coloum

尽管直接使用“Rings”可能会稍微简单一些，并使预测任务更接近原始数据结构。但是使用“Age”可能会使预测从生物学角度更容易解释，因为它们直接代表鲍鱼的年龄。因此我使用Age=Rings+1.5这个公式在数据集中添加名为Age的新的一列并且删除了原始的Rings列。

2.1.3 Dealing with Sex Columns

机器学习算法通常需要数字输入，Abalone数据集中的“性别”列包含分类值“F”（女性）、“M”（男性）和“I”（婴儿），需要先进行编码，然后才能在大多数机器学习模型中有效使用。

One-hot encoding converts each category value into a new binary column and assigns a 1 or 0 (True/False) value to those columns. For the 'Sex' column with three categories, it would create three new columns, one for each category ('F', 'M', 'I'), with binary values indicating the presence of each category.

Here I add three new columns called SexM, SexF, SexI and delete the Sex columns.

2.2 EDA

2.2.1 单一变量描述性统计分析

单个变量的描述性统计涉及总结和分析数据以描述鲍鱼数据集主要特征，该部分不对数据生成过程做出任何假设。在这个部分使用的是原始的数据集（abalone\_orgin）分析了集中趋势的度量（如平均值和中位数）、变异性的度量（如范围、方差和标准差）以及图形表示（如直方图）。

Summary+Histogram

2.2.2 变量间相关性分析

双变量分析可以清楚地了解每个特征在存在其他特征的情况下如何受到影响。它还帮助我们理解和识别显著特征，克服多重共线性效应和相互依赖性，从而提供对隐藏数据噪声模式的见解。

这个部分使用abalone\_Age\_DummySex数据集，其中，我们最关心的就是性别因素对于年龄预测的影响，通过相关性分析图表，可以观察到，三种性别与其他变量的关系是两条平行的直线，这暗示着性别对其他predictors以及age没有相关性，这个可以作为接下来建模的参考。不过相关性分析只是一个基础的分析方法，还不足以支持删除Sex变量。

ggpairs

2.2.3 最佳子集选择

鲍鱼数据集具有多个特征，直接将所有predictors用作模型输入并不是一个明智的选择，需要对输入变量进行筛选。由于鲍鱼数据集变量数目并不多，我们在这里选取最佳子集选择。在统计学和机器学习中，最佳子集选择是一种用于从一组预测变量中选择最佳预测子集的方法，以便构建模型。

我将leaps包中的regsubsets函数应用在abalone\_Age\_DummySex以及abalone\_Age\_NoSex两个数据集上，试图求出两个数据集的最佳子集。然后我得出如下结论：1. 在拥有Dummy变量的数据集abalone\_Age\_DummySex上，函数返回“1 linear dependencies found Warning”，这说明该数据集存在线性相关的变量，并不利于做回归预测。2. 在没有Sex的数据集上面，最佳子集的变量为6（7 total），并不包含Height，且在Adj Rsq，BIC，Cp图中表现一致。这两点结论为后面建模提供统计学支撑，但是依然不造成任何变量的删除。

三个图

Part 3 Modelling

在机器学习建模部分，报告中使用了Lp Norm线性回归以及随机森林算法。这个部分的核心方法论分为两步：1. 将数据集拆分成训练集和测试集（大概比例为8：2），再拆分时采用固定的随机数种子，确保多次训练和测试也能得到相同的结果；2. 通过定义函数，将机器学习在训练集上的训练，在测试集上的预测，使用RMSE计算模型误差以及最终可视化的过程封装成Pipeline（通过调用perform\_cv\_glmnet和perform\_random\_forest）即可快速实现对数据集的机器学习流程。

**数据集配置：**abalone\_Age, abalone\_Age\_Dummysex, abalone\_age\_NoSex, abalone\_age\_SexM, abalone\_age\_SexF, abalone\_age\_SexI及每个数据集对应被划分的训练集和测试集。使用这些数据集满足多种分析需求。

3.1 Lp Norm Linear Regression (Ridge, LASSO and Elastic Net)

3.1.1 Reason for choosing This instead of Decision Tree

线性回归模型（包括经过正则化处理的模型）通常比决策树模型更容易解释。在Lasso、Ridge和Elastic Net回归中，每个特征的权重可以直接解释为对目标变量的影响，这对于统计分析和结果解释非常重要。Lasso回归（L1正则化）特别适用于特征选择，因为它可以将不重要特征的系数缩减到0。如果鲍鱼数据集中有许多不重要的特征，Lasso回归可以自动进行特征选择并简化模型。如果鲍鱼数据集中的特征高度相关（多重共线性），Ridge回归（L2正则化）特别有用。Ridge回归通过添加一个正则化项来减少参数估计的方差，从而提高模型的稳定性和泛化能力。

3.1.2 Introduction to Lp Norm Linear Regression

Lp线性回归是一类通过在传统线性回归基础上引入正则化项来优化模型的方法。这些正则化技术旨在解决过拟合、多重共线性等问题，并在某些情况下实现特征选择。

三者的数学模型

3.1.3 Model Assumption

1. 线性关系假设

模型假设预测变量（如鲍鱼的长度、直径、高度、重量等）和响应变量（如年龄或环数）之间存在线性关系。这意味着响应变量可以表示为预测变量的加权和，加上一个误差项。

2. 多重共线性假设

虽然传统的线性回归模型假设预测变量之间没有完全的多重共线性，Lp线性回归可以容忍一定程度的多重共线性，并通过正则化减轻其负面影响。鲍鱼数据集中的预测变量高度相关（如整体重量和去壳重量），使用Ridge或Elastic Net可能更合适，因为这些方法可以处理这种多重共线性。

**3. 适当的alpha和lambda选择**

模型假设通过交叉验证选择的alpha和lambda值能够产生一个良好的正则化模型，既不过度拟合也不欠拟合。

**4. 特征空间的恰当表示**：模型假设输入特征已经被恰当地转换或选择，以便于线性模型能够捕捉到目标变量与特征之间的关系。

3.1.4 Design of Experiment (DoE)

在第三部分开头提到的为了满足多种分析需求设置的数据集共有5个，同时我设置了3个不同的alpha数值（0，0.5，1），分别对应着Ridge，Elastic Net和Lasso回归，理论上需要进行15次机器学习才能分析结果，但由于时间限制且15个结果分析需要消耗大量的精力。我设计了如下实验方法进行分析：第一部分是使用相同的数据集搭配不同的alpha数值来验证模型假设三；第二部分是使用不同的数据集搭配相同的alpha数值来验证模型假设四。这样就可以通过8次实验就可以得出结论。试验设计如下：

DOE Pic

3.1.5 Result Analysis

由于本部分的模型假设1和2在第二部分中已经被证实是正确的，下面我着重根据假设3和4进行分析，根据实验结果，发现如下：

1. 在相同数据集Abalone\_Age\_NoSex下，通过更换alpha的值，模型最终在训练集和测试集上的误差RMSE均处于2.1至2.2之间，随着alpha从0至1变化而变化，但是并不显著。
2. 照Sex不同的数据分成三组后，Infant数据集的RMSE远低于Abalone\_Age\_NoSex数据集，而SexM和SexF数据集的RMSE略高于该数据集，模型的性能变化较大，这可能反映了不同性别样本的内在差异。这意味着如果选择拆分后对SexI的预测效果更好。
3. 在第二部分实验中，alpha=1，随着lambda增大，模型最佳时只有6个变量，相关系数为0的变量是Height。这与Part 2中的Best Subset Selection结果相同。同时，不同数据集的最优lambda\_min值有所不同，这表明数据的不同特性可能需要不同程度的正则化。
4. 模型性能：从RMSE来看，Ridge回归（alpha=0）在这个特定的数据集上表现最好，尽管改进幅度非常小。弹性网络（alpha=0.5）和LASSO回归（alpha=1）的性能相似，但略微逊色。
5. 变量选择：LASSO回归通过将一些系数减至零来进行变量选择，而弹性网络和Ridge回归则保留了所有变量。这可能意味着在需要进行变量选择时，LASSO或弹性网络（当alpha接近1时）可能更为适用。
6. 正则化强度：lambda\_min的值随alpha的减小而减小，这表明对于Ridge回归，可能需要更强的正则化来达到最佳性能。

3.2 Random Forest

3.2.1 Reason for choosing This instead of Neural Network

由于随机森林是基于决策树的集成方法，随机森林提供了较好的模型解释性。我们可以容易地查看哪些特征对预测结果影响最大，甚至可以对单个决策树的决策路径进行解释。相比之下，神经网络通常被视为“黑盒”模型，难以解释其内部工作机制和决策过程。除此之外，不像神经网络那样对输入特征的缩放敏感。决策树基于特征的阈值进行分裂，因此不需要特征标准化或归一化。由于采用了多个决策树的平均预测，随机森林对于数据中的噪声和异常值通常更为鲁棒。

3.2.2 Introduction to Random Forest

随机森林是一种流行且强大的机器学习算法，属于集成学习方法的一种。它由多个决策树构成，通过对这些决策树的结果进行汇总来进行预测，从而提高模型的准确性和鲁棒性。

随机森林的工作原理：

自助采样（Bootstrapping）：随机森林通过对原始数据集进行多次自助采样（有放回的随机抽样）来生成多个不同的训练子集。

建立决策树：对于每个训练子集，算法构建一个决策树。在构建决策树的过程中，每次分裂节点时，不是从所有特征中选择最优特征，而是从特征的随机子集中选择最优特征。这种随机性的引入有助于提高模型的多样性，减少过拟合。

汇总预测：

对于分类任务，随机森林通过多数投票的方式来确定最终的类别，即每个决策树给出一个预测结果，类别获得最多“票数”的为最终预测结果。

对于回归任务，随机森林将所有决策树的预测结果取平均值作为最终的预测结果。

数学模型

3.2.3 Model Assumption

变量相关性：随机森林通过在每个决策树分裂时从特征的随机子集中选择最优特征，减少了变量间高相关性的影响。因此，模型假设即使特征之间存在一定程度的相关性，也不会对模型性能产生显著的负面影响。

模型复杂性与过拟合：随机森林通过集成多个决策树来提高预测准确性，同时通过引入随机性减少过拟合风险。因此，模型假设即使是在高维数据或复杂的数据结构上也能保持良好的泛化能力。

特征重要性：随机森林可以评估各个特征对预测结果的贡献度，因此模型假设某些特征可能比其他特征对预测年龄更重要。

无需特征缩放：随机森林对特征的尺度不敏感，因此模型假设不需要对特征进行标准化或归一化处理。

3.2.4 Model Tuning

在R语言中使用randomForest函数优化模型性能时，关键参数调整包括ntree（树的数量）、mtry（分裂时考虑的特征数量）、nodesize（终端节点的最小样本数）、sampsize（每棵树的样本大小）和maxnodes（每棵树的最大节点数）。增加ntree可以提升模型稳定性，调整mtry影响模型多样性和灵活性，而nodesize和maxnodes的调整有助于防止过拟合。

除了randomForest内部的参数调整之外，在对Age预测的时候还可以多次尝试不同的特征组合，比如移除Sex或是Height，查看模型的方差解释性和RMSE变化。

3.2.5 Result Analysis

在这个部分，我使用了四个数据模型对随机森林进行训练。第一个模型：使用所有特征；第二个模型：排除性别特征；第三个模型：排除身高特征；第四个模型：排除性别和身高特征。这四个模型中我都默认使用了树的数量500和每次分裂考虑的变量数2，得到如下结论：

特征的影响：排除特定特征（如性别或身高）对模型整体性能的影响相对有限。这可能表明这些特征对于预测年龄不是决定性因素，或者它们的信息可能已经间接地通过其他特征被模型捕获。

训练误差与测试误差：在所有模型中，训练误差 consistently 显著低于测试误差。这是过拟合的一个典型迹象，表明模型可能对训练数据过于敏感，捕捉到了一些训练数据中的噪声，而这些噪声在测试数据中并不普遍存在。

模型解释的方差：所有模型都能解释大约55%左右的方差，这表明随机森林模型在捕获数据集内在结构方面具有一定的能力，但仍有一定比例的方差未被模型解释，可能由数据中的随机噪声或未观察到的变量引起。

平均平方残差：平均平方残差在各个模型之间的差异不大，这进一步证实了移除单个特征对模型预测性能的影响有限。

运行时间：各个模型的运行时间相对一致，表明特征的移除对计算效率的影响不大。这可能是因为随机森林算法本身在处理特征时具有较高的效率，且模型中树的数量（500棵）是决定运行时间的主要因素。

Part 4 Model Comparison

Lp线性回归结果概述：

模型系数：Ridge回归为每个特征提供了一个系数，这有助于理解每个特征对预测变量的影响。

训练误差（RMSE）：在各个模型中变化，从1.55983到2.541056不等。

测试误差（RMSE）：从1.646548到2.502254不等，显示了模型在测试数据上的泛化能力。

运行时间：非常快，仅需约0.08到0.14秒。

随机森林结果概述：

模型细节：每个模型均使用了500棵树，每次分裂考虑的变量数为2。

平均平方残差：在不同模型中略有变化，但基本维持在4.717192到4.764544的范围内。

解释的方差百分比：大约在54.83%到55.27%之间，表明模型能够解释一半以上的数据方差。

训练误差（RMSE）：从1.012617到1.026651不等。

测试误差（RMSE）：从2.064943到2.08656不等，显示了模型在测试数据上的泛化能力。

运行时间：相对较长，需要约4.35到4.51秒。

比较：

性能：随机森林和Ridge回归的测试误差相近，表明两种方法在这个特定数据集上有相似的预测能力。不过，随机森林在解释方差方面略占优势。

解释性：Ridge回归提供了每个特征的系数，从而提供了更好的模型解释性。随机森林虽然可以评估特征重要性，但整体解释性不如线性模型。

计算效率：Lp norm线性回归的运行时间远低于随机森林，特别是在处理大型数据集时，这一点尤其重要。

对非线性关系的处理：随机森林能够更好地处理非线性关系，而Ridge回归可能需要额外的特征工程来捕捉这些关系。

结论：

选择哪种模型取决于具体的应用场景。如果模型解释性是关键考虑因素，并且数据关系接近线性，Ridge回归可能是更好的选择。而对于存在复杂非线性关系的数据集，或者当模型性能是唯一关注点时，随机森林可能提供更好的结果，尽管以较长的训练时间为代价。在实际应用中，最佳做法是根据数据的特性和业务需求对多个模型进行评估和比较。

**Reference**

[Abalone Dataset (kaggle.com)](https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset)

**Appendix**

是多少