**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

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 + VisceraWeight + 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

Part 3 Modelling

Part 4 Model Comparison

**Reference**

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