Predicting Abalone Age Using Regression Models - Proposal

Team: Rakesh Kumar Busa, Jeremy Driscoll, Daniel Ackom

Introduction

Determining the age of abalones is a crucial aspect of marine biology and fisheries management, as it provides insights into growth patterns, population dynamics, and sustainability measures. Traditional methods for age estimation involve physically counting growth rings on an abalone's shell, a process that is both time-consuming and prone to human error. This study aims to address this challenge by investigating the predictive capability of various regression models to estimate abalone age based on physical attributes.

Research Question

Specifically, we seek to answer the question: *Which regression model provides the most accurate and interpretable predictions of abalone age while addressing challenges such as multicollinearity and non-linearity?* By systematically evaluating different regression models, we aim to identify the most effective approach for predicting abalone age while balancing accuracy and interpretability.

Dataset Description

This study utilizes a dataset from the UCI Machine Learning Repository, containing 4,177 observations of abalones. Each observation is characterized by numerical attributes, including length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight. The dataset also includes the number of rings, which serves as a proxy for age, with the actual age calculated as the number of rings plus 1.5 years. A categorical variable representing sex (male, female) is also included. The dataset presents challenges such as multicollinearity among weight-related variables and potential non-linearity in the relationship between predictors and age, necessitating careful model selection for robust predictions.

Methodology

The methodology begins with preprocessing, addressing missing values, encoding categorical variables, and normalizing numerical features for consistent scale. A variety of regression models are applied to predict age. Ordinary Least Squares (OLS) regression serves as a baseline, while Ridge regression, which incorporates L2 regularization, is tested to mitigate multicollinearity. Lasso regression is employed to assess the impact of feature selection by shrinking insignificant coefficients, and polynomial and spline regressions are explored to capture non-linear relationships.

Model Evaluation

Model performance will be evaluated using statistical metrics such as Mean Squared Error (MSE), Residual Sum of Squares (RSS), and R-squared scores.

Justification of Model Selection

Model selection is guided by statistical considerations and dataset characteristics. OLS regression provides a benchmark but is vulnerable to multicollinearity. Ridge and Lasso regression address this by regularizing coefficients, with Ridge preserving all features and Lasso performing feature selection. Polynomial and spline regressions are tested to capture potential non-linear relationships. The comparative analysis of these models seeks to balance prediction accuracy and interpretability, ensuring the selected model generalizes well to unseen data.