Abalone Age Classification

Capstone Project

A Project to develop a model for Abalone age Prediction

By W.A.R.H. Rodrigo

ABSTRACT

This report take aim at accurate prediction of abalone age from its physical measurements. The Abalone dataset from the UCI-Machine Learning Repository consisting of 9 attributes and 4177 observations was subjected to various methods of model fitting and comparison of performances. The descriptive analysis conducted, revealed that there is a specific difference in physical attributes of the 'infant' and 'non-infant' abalone. In the advanced analysis, the abalone age prediction was initially conducted for the entire data set using statistical and Machine learning classification methods. The performances of the models are compared on the basis of accuracy and f1-score. Logistic Ridge regression with 87.5% accuracy as the best model for abalone age prediction respectively. Furthermore, Random Forest regression model was introduced for prediction of the numerical age of abalone considering the usefulness of the numerical age in price formulation.

1. Introduction

Abalone, characterized by their single shell and distinctive appearance, resemble oysters and mussels more than typical snails. Inhabiting the cold coastal waters of every continent, these marine snails belong to the family Haliotidae, encompassing seven West Coast species and approximately 60-100 species globally. Culturally, abalone holds significance for Native tribes on the West Coast, serving as a source of both meat and ornamental shells. Ecologically, abalones play a crucial role in habitat maintenance by regulating algal density. The vibrant iridescence of their inner shell layer makes abalone shells prized for decorative objects and jewelry, contributing to their appeal and economic value.

Beyond cultural and ecological importance, abalone holds culinary prestige worldwide. Recognized for its unique flavor, soft texture, and nutritional richness (high in protein, iodine, and selenium, offering omega-3), abalone is considered a luxury seafood. In Southeast Asia, it is particularly esteemed, with consumers willing to pay a premium for high-quality specimens. The scarcity of abalone, coupled with its slow growth rate (1 inch per year) and labor-intensive harvesting, contributes to its high market price. Given the economic relevance tied to abalone age, predicting age through physical measurements becomes imperative for farmers and distributors, aiding in market pricing decisions and enhancing the sustainability of the abalone industry.

2. Problem

Abalone is a flavorful yet expensive shellfish highly valued in cuisines around the world. By counting the number of layers in its shell, one can determine the age of an abalone. It involves cutting a sample of the shell, staining it, and counting the number of rings through a microscope, which can be a tedious process. Figuring out the price of an abalone is of utmost importance for farmers and sellers, as the price variation is age dependent. Hence the prediction of age by considering the physical measurements which are easier to obtain is our main objective of this project.

In order to achieve the main objective, it is required to identify the major physical measurements that contribute for accurately predicting the abalone. Hence it is divided to two sub categories as,

- Physical measures related to size
- Physical measures related to weight

3. Data and Methodology

The Abalone dataset consisting of 9 variables and 4177 observations was extracted from UCI Machine Learning Repository which is sourced from an original (non-machine-learning) study: named "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait". By the

authors, Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford, consists of lab is measurements of blacklip abalone collected from the Bass Strait, which separates Tasmania from mainland Australia *(Nash et al., 1994).*

Source: https://archive.ics.uci.edu/ml/datasets/abalone

Table 3.1 Variable Description

Variable	Type	Description
Sex	Nominal	Male, female, infant
Length	Continuous	Longest shell measurement (dm)
Diameter	Continuous	Measurement perpendicular to length (in dm)
Height	Continuous	with meat in shell (in dm)
Whole weight	Continuous	the whole abalone weight (in grams)
Shucked weight	Continuous	weight of the meat in abalone (in hg)
Viscera weight	Continuous	gut weight after bleeding (in hg)
Shell weight	Continuous	weight after being dried (in hg)
Rings	Discrete	+ 1.5 gives the age in years

Methodology

Initially the dataset was subjected to initial preprocessing. Next the description analysis was conducted in order to answer the questions identified in the objective. Next the model training was conducted initially for the categorical age prediction. Model training conducted for few models and the best model was chosen in the basis of f1 score and accuracy. Later on the sampling techniques were used to improve performance of the model, but it was not a success,. However the need for a regression model was required as the age value prediction was important when it comes to determining the price of an abalone. Hence Regression models were fitted as well. The data profuct/application allows the use to predict the age as well as the Age category it falls to.

4. Results

Model training for Classification variables gave the below output and the best model was identified to be Logistic Ridge regression. However sampling techniques were applied to improve model performance but they were unsuccessful.

Model	Accuracy (%)	F1-Score (weighted)
Logistic Ridge	87.54	0.86
Logistic Lasso	87.42	0.85
Logistic Elastic-net	87.42	0.85
KNN	86.71	0.84
Random Forest	86.82	0.85
Gradient Boosting	86.71	0.85

Model training for Regression was conducted for Abalone age as a numerical value. Results of those models are provided as below. Random forest regressor was identified as the best performer as it has the lowest RMSE and the largest R square value.

Regression Model	RMSE	R square
Linear Regression	2.222	0.521
Ridge Regression	2.250	0.509
Lasso	3.214	-0.001
E-net	3.084	0.078
KNN	2.349	0.465
CART Regression	2.868	0.202
Random Forest Regression	2.186	0.536
Gradient Boost Regression	2.196	0.532

5. Discussion and Conclusions

The main aim of this project is the accurate prediction of abalone age from its physical measurements. The descriptive analysis was conducted to get an insight about the data and how the physical measures contribute in age determination of abalones. The variables, whole weight, length, diameter and height was found to be significant in determining the age of old abalones. It was found that there is a significant difference in factors that affect the age of abalones in the two sexes of abalones infant and non- infant. According to the analysis it was because the Infants were underdeveloped which led to the difference in age determining features. Most interesting finding is that the abalones show both allometric and isometric growth for certain physical measures.

The advanced analysis was conducted in order to achieve the goal of developing a classification model for the accurate prediction of all age types of abalone. However, priority was given to the identification of old abalones as they demand high prices. Because inaccurate predictions may incur losses to the stakeholders in the industry. Initially, Machine learning models and statistical models were fitted to the entire data set, where the performances of the models were conducted on the basis of accuracy, and individual f1-scores. The best model; Logistic Ridge regression (87.5%) was insufficient in accurate prediction of individual categories. Sampling techniques was implemented as a remedy, but was unsuccessful. Taking into account the differences exhibited by the two sexes, infant and non-infants, separate models could be fitted to the two groups as an improvement. However in the current project it was not attempted.

The previous predictions of abalone age in kaggle kernals, rpubs, towardsdatascience, github were included in the range of 26%-55% accuracies for modeling all 29 classes (target: 'Rings' variable). However, for modelling for 3 classes, accuracies were in the range of 79%-93%. Hence our results can be justified as they fall within the expected ranges. Furthermore, a regression model was trained to predict the numerical age of abalones to reduce the effort required by sellers and buyers of abalone because it will be easier for them to name prices directly from the numerical age. **Random Forest model** with low RMSE (Root Mean Square Error) of 2.186 was developed for Prediction of age. However, the model performance can be further improved considering the expected accuracies. Implementation of Neural Networking and models relevant to deep learning are suggested for improving the accuracies. To overcome the limitations in the data and model training, its suggested to obtain a larger dataset with balanced 'Age' classes.