ABALONE AGE PREDICTION MODEL

Submitted by:

Abhilekh Verma

Contents

ACKNOWLEDGMENT 3

INTRODUCTION 4

About Abalone 4

Problem Statement 4

Objective 4

Data Description 4

Methodology 5

Model Building 5

Identification of possible problem-solving approaches 6

Testing of Identified Approaches (Algorithms) 6

Deployment 6

- All the models are performing well after cross validation Hyper parameter Tuning . 7
- We can use Linear Regression for our model. 7

References 7

Learning Outcomes of the Study in respect of Data Science 7

ACKNOWLEDGMENT

I am thankful to Data Trained for the guidance. They have covered the topics like Machine Language, Python & SQL. There are also some references taken from YouTube & google search.

INTRODUCTION

About Abalone

- Abalone are marine snails.
- The shells of abalones have a low, open spiral structure, and are characterized by several open respiratory pores in a row near the shell's outer edge. The thick inner layer of the shell is composed of nacre (mother-of-pearl), which in many species is highly iridescent, giving rise to a range of strong, changeable colours, which make the shells attractive to humans as decorative objects, jewellery, and as a source of colourful mother-of-pearl.
- Abalone can live up to 50 years, depending on a species.
- The speed of their growth is primarily determined by environmental factors related to water flow and wave activity.
- Abalone vary in size from 20 mm (0.8 in) (Haliotis pulcherrima) to 200 mm (8 in) while Haliotis rufescent is the largest of the genus at 12 in (30 cm).

Problem Statement

- The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.). Further information, such as weather patterns and location (hence food availability) may be required to solve the problem).

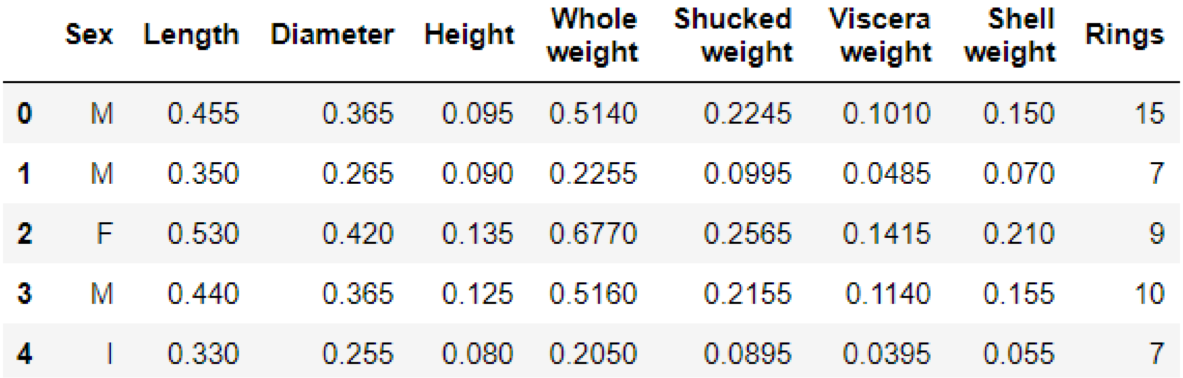
Objective

- Objective of the project is to predict the age of the abalone using machine learning algo.
- We will use physical characteristic of the abalone to predict age using learning algo.
- We will use Machine Learning algorithms like Linear Regression, Lasso, Ridge, Elastic net to analyse all the parameter.

Data Description

- The source of data is taken from GitHub.
- 8 physical characteristics
- Sex
- Length
- Diameter
- Height
- Whole weight
- Shucked weight
- Viscera weight
- Shell weight

Data sample looks using panda.



	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

Methodology

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the age of abalone because it gives the best outcome for the assurance of age of abalone.
- The dependent variable is Rings, whereas other variables i.e., Length, Height etc. are independent variables.
- While handling the effectiveness of the data model, various types of errors have occurred like over fitting, introduced from having too large of a training set and bias occur due to small of a test set.

Model Building

a. Linear Regression, where we got 100% accuracy.
b. Lasso, where we got 98% accuracy.
c. Ridge, where we got 99% accuracy.
d. Elastic Net, where we got 98% accuracy.

Hardware and Software Requirements and Tools Used

a. Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram
b. Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

Identification of possible problem-solving approaches

Following models are used for solving the problem:

a. R2 score: R 2 score is used to evaluate the performance of a linear regression model.
b. Linear Regression: Linear regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results).
c. Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization
d. Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model
e. Elastic Net: Elastic-Net is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
f. Random Forest Regressor: is a meta estimator that fits a number of classifying decision tree on various sub samples of the dataset.
g. Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
h. Z-score: For checking and removal outliers in the dataset.
i. Datasets: overall data.
j. Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.

Following statistical and analytical approach followed:

a. regression coefficients are marginal results.
b. Started with univariate descriptive and graphs.
c. bivariate descriptive, again including graphs.
d. Model building and interpreting results.

Testing of Identified Approaches (Algorithms)

a. Train Test Split
b. Linear Regression
c. Lasso
d. Ridge
e. Elastic net
f. Random Forest Regressor
g. Cross Validation
h. Hyper Parameter Tuning Using Grid Search CV.

Key Metrics for success in solving problem under consideration.

1. Analysed data for any outliers and removed it by z-score method.
2. Analysed data for any skewness.
3. Handling class imbalance problem by oversampling the minority class.
4. Cross Validation for cross validates the accuracy-score from overfitting.
5. Hyper parameter tuning using Grid Search Cv for making the prediction better.

Conclusion

- All the models are performing well after cross validation Hyper parameter Tuning.
- We can use Linear Regression for our model.

References

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.

Learning Outcomes of the Study in respect of Data Science

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc.

It also helped me to learn how to apply various model techniques on data and enable predications.