Data processing and linear regression Report
Conducting Linear regression on the Abalone dataset

Overview of project:

The age of an Abalone may be determined by cutting through its shell, staining it and then counting the number of rings that appear under a microscope. By multiplying the number of rings by 1.5 we can obtain an accurate measure of the Abalones age. However, this task is fairly time consuming and there may be other physical characteristics that offer comparably accurate age estimates for less effort. Therefore, the question of interest is whether there exists other physical features that provide accurate insight into estimating an Abalones age.

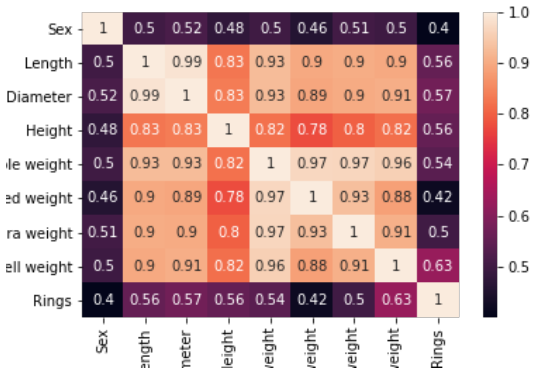
To answer this question the Abalone Dataset[1] was utilized. The dataset provided 4177 unique entries physical properties of Abalones including:

- Sex – Male, female, or infant;
- Length - Longest shell measurement (mm)
- Diameter - Perpendicular to length (mm)
- Height - with meat in shell (mm)
- Whole weight - whole abalone (grams)
- Shucked weight - weight of meat (grams)
- Viscera weight - gut weight after bleeding (grams)
- Shell weight – Measured after being dried (grams)
- Rings - gives the age of abalone

Hence, this project aimed to identify two possible physical measures of the abalone that would aid in less effort to determine the animal's age.

Feature selection:

After completing pre-processing of data it was ready for analysis. The first step of analysis involves assessing the correlation matrix. This matrix provides insight into what variables highly correlate with our variable of interest, number of rings (fig below). Scores are between -1 and 1 with values closer to -1 or 1 indicating a strong relationship and scores around 0 indicating no relationship.

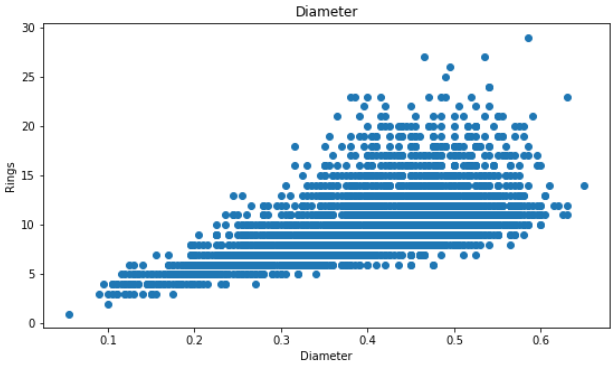


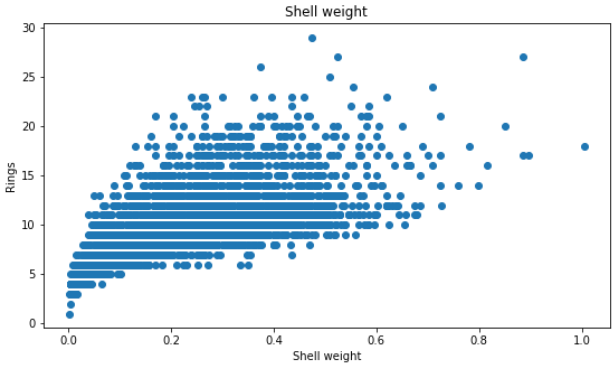
[1] Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994), "The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288). https://archive.ics.uci.edu/ml/datasets/abalone

The first thing of notice is that all of the physical features positively correlate with number of rings. As we had no negative correlation with rings two positive features were chosen for further analysis. These being the two most significant features:

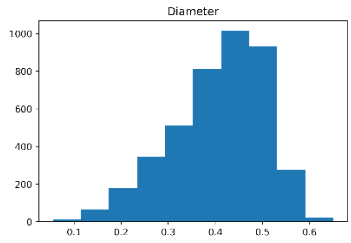
- Weight (0.63); and
- Diameter (0.57)

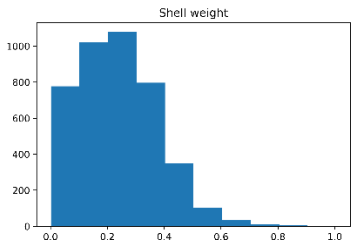
Next, we looked deeper into these features relationship with Abalone rings by plotting weight and Diameter against number of rings. This was conducted using a scatter plot as a general relationship can be derived by inspection. Clearly the graphs below show that as diameter and shell weight increase it is a reasonable assumption to expect an increase in the number of rings of the Abalone. Hence, these features offer a good starting point for deriving a model that will estimate the age of the Abalone

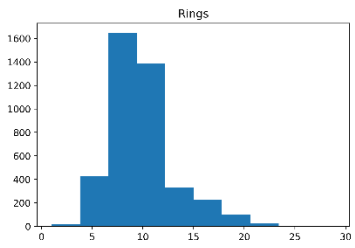




We then looked at the distribution of our parameters of interest. By looking at the distribution we can observe features such as where our data is centred and its shape. By observing the graphs below we can see that all of these features are skewed, with shell weight and rings skewed to the right and diameter skewed to the left. This is an interesting finding as its not an unreasonable assumption that with a greater diameter the Abalone will have a bigger shell and hence greater shell weight, however, this doesn't appear to be the case here. The data shows that the average Abalone has around 10 rings (lives for 15 years), growing up to 0.4mm in diameter with a shell weight of around 0.3 grams.







Evaluating our model

Results showed that when using two feature variables and normalising our data in over 30 experiments it was found that the mean-square error (how far data points deviated from our modelled line) was around 0.74. However, when assessing r-squared scores (a percentage between 0-100% indicating how much variance our model captures) our model was only able to capture around 0.40% of rings variation.

To assess if there was a better model available linear regression with all physical features was ran in order to see if this would produce a better model. This experiment was conducted with equal

conditions to our first experiment with the only difference being the number of variables used within our model. Firstly, whilst normalising the data reduced the mean square error it didn't significantly affect the r-squared score with both models explaining 52% of variance within our target variable. Secondly, we can observe that this explained ring number variation 12% better than our simpler model. Whilst this is an increase in model accuracy it also requires an increase in effort as more physical characteristics of the abalone are required to be measured to get this improved estimation. If one was to measure all these variables it may have been less effort to stain the shell and count the rings. The finding from this experiment indicates that measuring the physical properties of the abalone provide a very rough estimate of the age of the creature.

References:

Code Resources:

1. Correlation Matrix, Scatter Plot, Linear Regression Code: "Linear Regression on Boston Housing Dataset"< https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-f409b7e4a155.
2. Histogram Code by Jake Warby "Exercise 1.3 Solution in Python", https://edstem.org/au/courses/6212/lessons/13871/slides/110007.
3. Data Cleaning with numpy Code by Unknown Author "Exercise 1.4 Part1 Solution", https://edstem.org/au/courses/6212/lessons/13871/slides/111793.