31/03/2021
Nikhil Khetani

Abalone Dataset Exploration and Cleaning Review

The data is taken from the abalone dataset which can be downloaded from https://archive.ics.uci.edu/ml/datasets/abalone. The data describes various attributes of them for use in training machine learning models. Features include sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and rings. The number of rings can also be used as a target variable in regression models.

1 Introduction

An abalone is a large marine gastropod mollusk – sea snail. It is often considered a culinary delicacy and is among the worlds most expensive seafood. However, this popularity has led to overfishing and it is illegal to collect wild abalone from the sea in many parts of the world.

2 Data Exploration

2.1 Initial plan

In order to ensure data quality, data would first be tested for missing values, then graphs would be plotted in order to ascertain the form of the data and identify outliers. The dataset has 4177 entries and 9 variables.

2.2 Data cleaning and Feature Engineering/ Key findings and insights

Firstly, data was tested for missing values or N/A values. After loading and testing the data, none were identified.

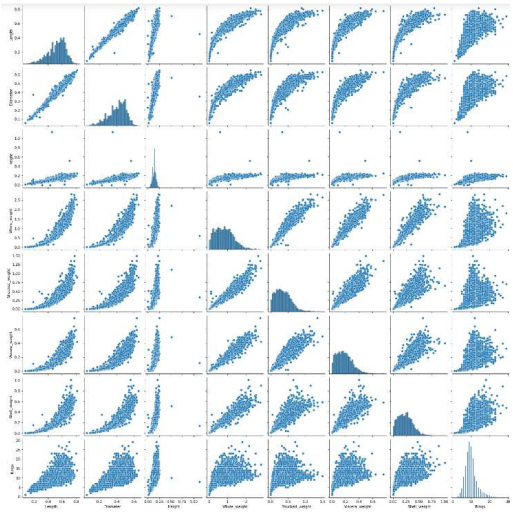

Figure 1: Pairplot of all features within data

Following this, a pairplot was created as seen in Figure 1 so that the data including its distributions and correlations could be more easily understood.

From the pair-plot it is clear length and diameter are skewed to the left whilst the various weights and the number of rings are skewed towards the right.

To make the data closer to a gaussian/normal distribution, apply exponential function to length and diameter, and apply logarithmic function to the weights and ring number. We can also identify two outliers in the height feature and will look at these in more detail first. To confirm these outliers, we will first view the top ten largest heights as shown in Table 1.

# Largest	Index	Height
1	2051	1.130
2	1417	0.515
3	1428	0.250
4	1763	0.250
5	2179	0.250
6	277	0.240
7	307	0.240
8	1528	0.240
9	2161	0.240
10	506	0.235

Table 1: Top ten tallest Abalones

The first two abalones in the table were identified as outliers (index 2051 and index 1417) and were removed from the dataset.

In order to convert the categorical data in numerical data, one-hot encoding was applied to the sex column, splitting it into three features: sex_F, sex_I and sex_M.

As the data was clearly skewed, the next step was to transform the data distribution to a more

gaussian shape. This was done by taking the exponential of the data:

$$T(x) = e^x$$

The effect of this transformation on the length feature can be seen in Figure 2.

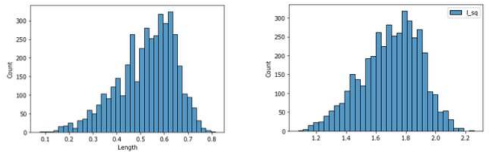

Figure 2: Histogram of length variable(left); histogram of exp(length) variable (right).

In addition to this, the data was scaled using the function:

$$T(x) = (x - x_{min})/(x_{max} - x_{min})$$

Min-Max scaling was used as outliers were previously removed from the dataset.

3 Hypothesis Testing

Three hypothesis were identified:

1) The transformed feature for length is normally distributed

2) Length and height are correlated

3) Length and weight are correlated

For the first of the three a formal significance test was conducted.

Null Hypothesis: Transformed feature for length is not normally distributed.

Alternative Hypothesis: Transformed feature for length is slightly normally distributed.

In order to be seen as significant, the maximum p-value accepted was set as 0.01.

After testing the normal distribution, the chi-squared statistic was calculated to be 242.41 and the p-value 2.29 *10^-53. As p<0.01, the hypothesis is accepted, and we can say that this feature in normally distributed.

4 Next Steps

The next steps for this dataset include training a model so that predictions can be made using this data.

5 Conclusion

Overall, this dataset is of a high quality and as far as can be observed, additional data is not needed immediately. The data is ready to be used for training models with non-numerical data converted to numerical data and all data scaled