Assignment 1 Report

Neha Devadas (z5332381)

Introduction:

In this report we aim to predict the Ring-age of Abalones based on their 8 features namely Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera Weight and Shell weight. For the experiment we are using a Linear Regression model with multiple input features. We compare the results of using all features, and just two selected features. We also investigate the effect of normalisation of data in each of the cases. A series of 30 experiments were conducted on each of the 4 cases and the metrics calculated for comparison are Mean RMSE, Std RMSE, Mean R2 score and Std R2 score.

Data Processing:

Q1. Data cleaning is performed on the column of 'Sex' attribute to replace values M, F, I (Male, Female and Infant) into numeric values 0, 1 and -1 respectively.

Q2. **Correlation Matrix:**

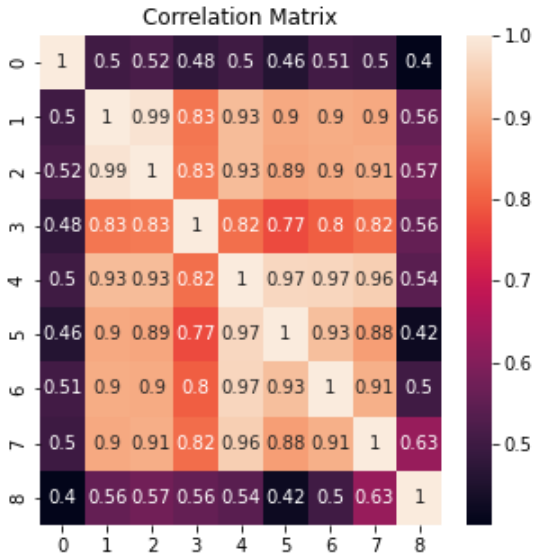

Figure 1: Correlation Matrix of data

The correlation matrix in the above figure was obtained by using a heatmap plot from the seaborn library. As expected, the diagonal of the matrix entirely consists of 1's as any variable's correlation with itself is 1. Normally, the values in the correlation matrix range from -1 to 1. It is interesting to note that there are no negative values in this matrix and the lowest value is 0.4.

Column 8 indicates the target variable i.e., Ring-age, and we are interested to see its correlation with the other 8 variables (features) of the dataset.

On further inspection, it appears that the highest positive correlation value is given by row 7 i.e., Shell weight. The value is indicated in the above diagram as 0.63.

The second highest positive correlation observed in the above figure is 0.57 given by row 2 i.e., Diameter.

For this experiment we are aiming to choose 2 of the features that give the highest positive correlation value. Hence, we select Diameter and Shell Weight as the 2 features for the model.

Feature 1: Diameter

Feature 2: Shell Weight

Target: Ring-age

Q3. **Scatter Plots:**

<u>Feature 1: Diameter</u>

The first feature that we have selected is the Diameter of the Abalone.

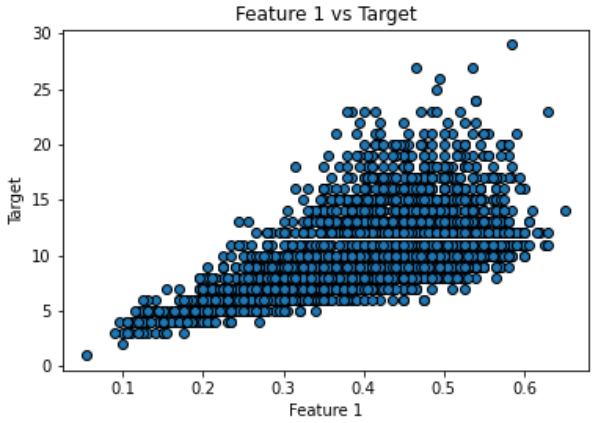

Figure 2: Diameter vs Ring-age

Figure 2 above shows the scatter plot of the Diameter vs Ring-age of the Abalone. On observing the graph, we can see that the trend is diagonally upwards confirming that it is a positive correlation. We can also see that most of the data is between 0.1 and 0.6 along with a few outliers.

<u>Feature 2: Shell weight</u>

The second feature that we have selected is the Shell weight. Similar to the previous feature, we can clearly observe in the below scatter plot (Figure 3) that the trend is moving upwards and towards the right indicating that it is a positive correlation.

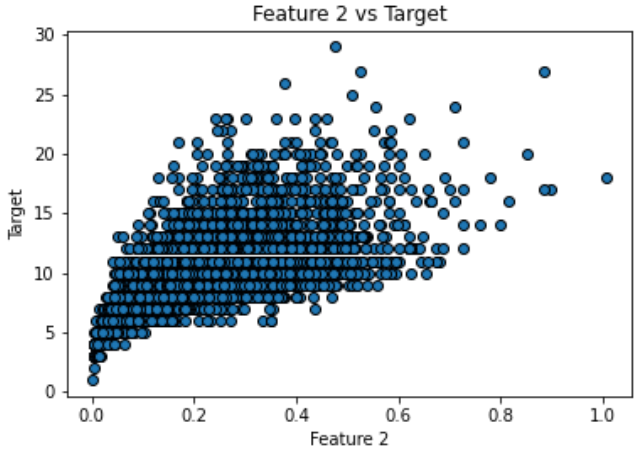

Figure 3: Shell weight vs Ring-age

We also note that the scatter plot is more densely packed between 0 to 0.5 indicating that most of the Abalone's Shell weight is between this range. We also see that as the shell weight increases beyond 0.6, it corresponds to a higher Ring-age value.

Q4. **Histograms:**

We have then plotted histograms of the 2 selected features as well as the target ring-age.

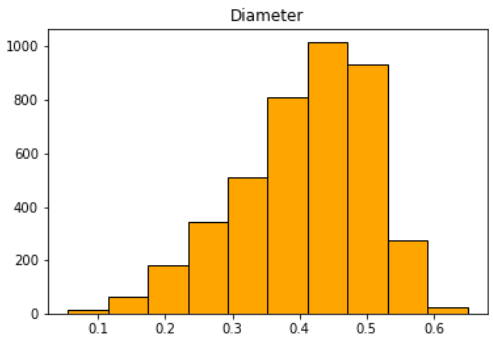

Figure 4: Histogram Distribution of Diameter

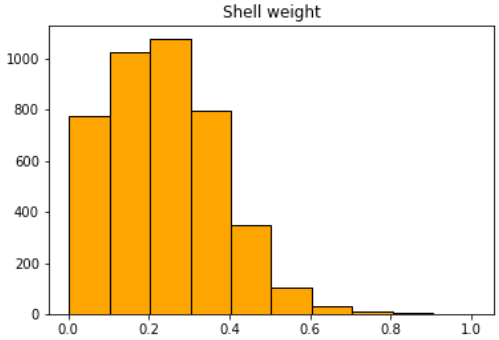

Figure 5: Histogram Distribution of Shell weight

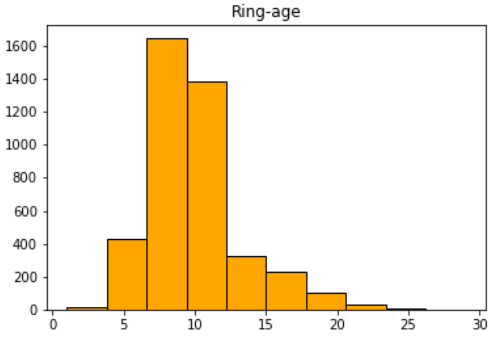

Figure 6: Histogram Distribution of Ring-age

We observe that the for the first feature in Figure 4, the distribution is unimodal and has 1 cluster. It is skewed slightly to the left as it trails off towards the left. The overall range is around 0.7 mm.

The second feature, the Shell weight, also has a unimodal distribution and is skewed more towards the right as we can observe that it trails off towards the right as observed in Figure 5. The overall range is about 0.9 grams.

Finally, on observing the ring-age distribution, we can see that it is unimodal and has a single cluster. It is not exactly symmetric as it is skewed slightly towards the right as seen in Figure 6. We also see that most of the data is between 5 to 15, but there are some outliers.

Q5. **Train-Test split:**

The data is then split into 60:40 train-test ratio randomly based on the experiment number. A series of 30 experiments is conducted and the experiment number is fed as the random seed to the split function.

Q6. **Data Visualisation:**

Next, we have plotted the distribution of all the different features and the target:

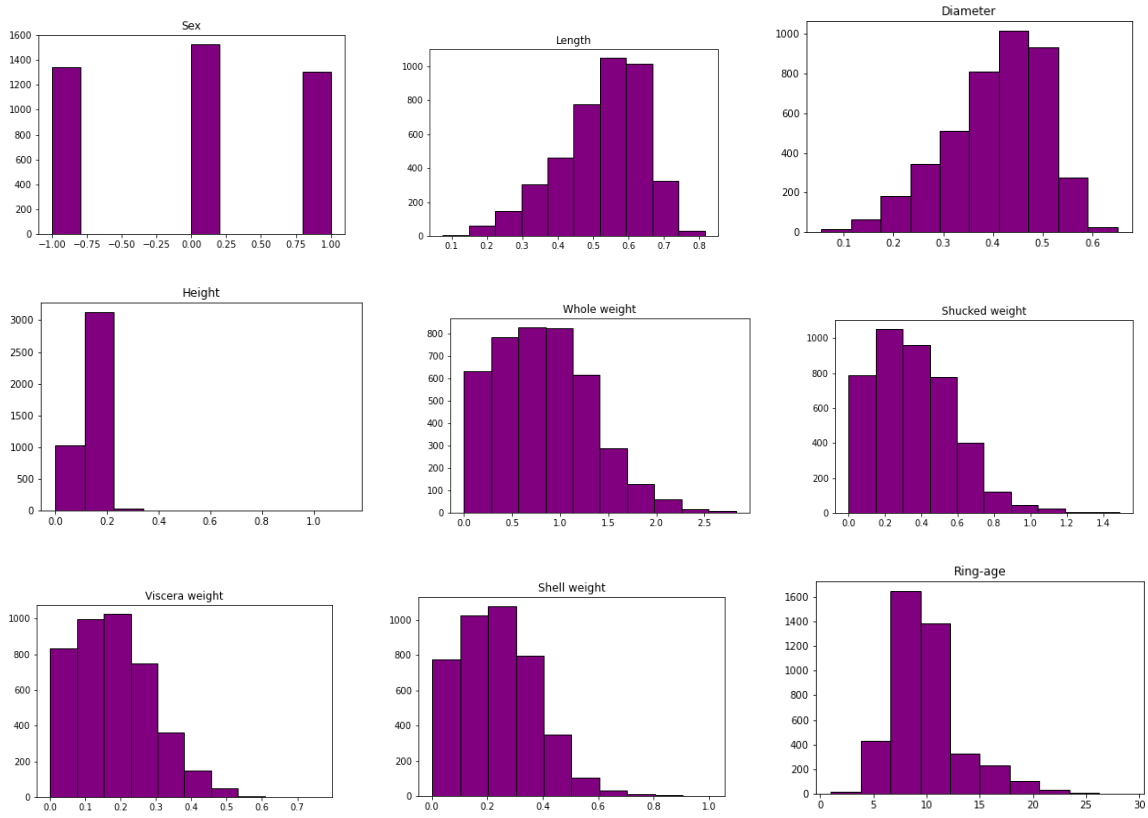

Figure 7: Histogram Distribution of all the columns in the data set

A general observation of the distribution is that for Sex there are 3 distinct values out of which there are a larger number of males. The height distribution is also limited to 3 distinct values. For Length and Diameter, the distribution is unimodal and skewed to the left. Whole weight, Shucked weight, Viscera weight and Shell weight also have single clusters each, are unimodal and skewed to the right. The target Ring-age also has a distribution slightly skewed to the right.

Modelling:

Q1. Q2. Q3. Q4.

For our Linear Regression model, we will be experimenting with 4 different cases: Using all input features with and without normalisation and using the best 2 features with and without normalisation. The 2 features selected here are the Diameter and Shell weight as mentioned in the Data Processing section of the report. The train-test split ratio here is 60:40.

Case 1: Using all features without normalising input data:

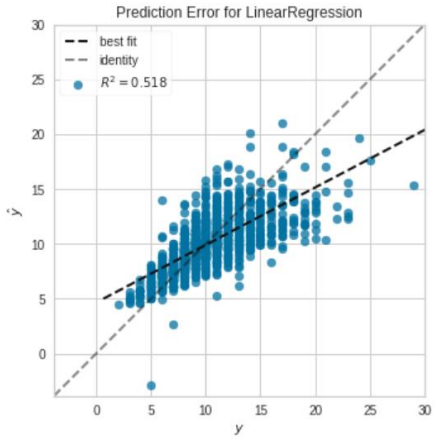

Figure 8.1: Prediction Error for Case 1

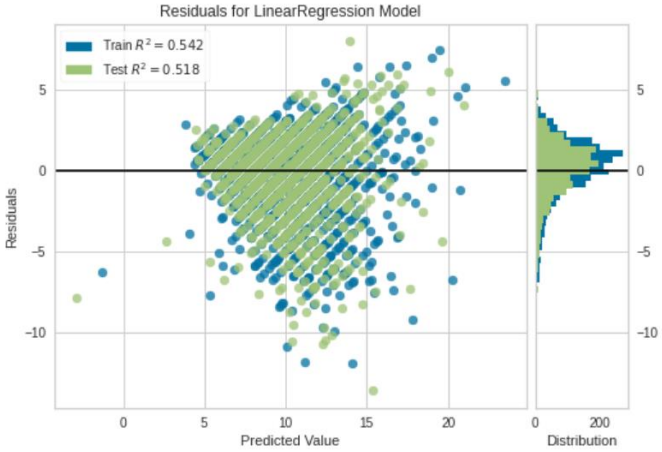

Figure 8.2: Residuals for Case 1

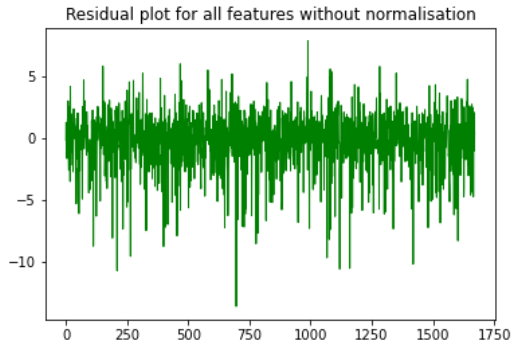

Figure 8.2: Residual plot for Case 1

After performing a single experiment, we observed that the RMSE score was 2.23010278 and the R2 score was 0.51740954. This indicates that the model using all features without normalisation explains only around 52% of the variation in the target variable. In Figure 8.1 we can clearly observe the prediction error of the model in Case 1.

After performing 30 experiments, we observe that our model has a mean RMSE value of 2.237461 and std RMSE value of 0.0512357. It also has produced a Mean R2 score of 0.5193237 and Std R2 score of 0.0224726801.

Case 2: Using all features with normalising input data:

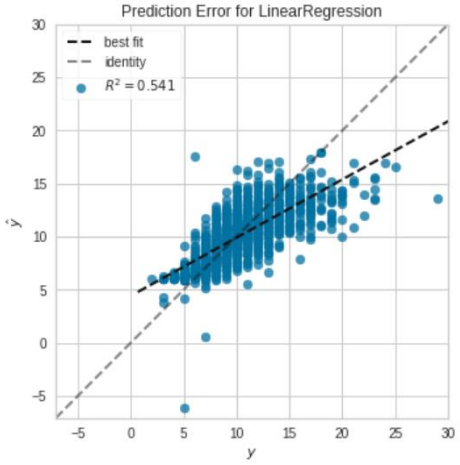

Figure 9.1: Prediction Error for Case 2

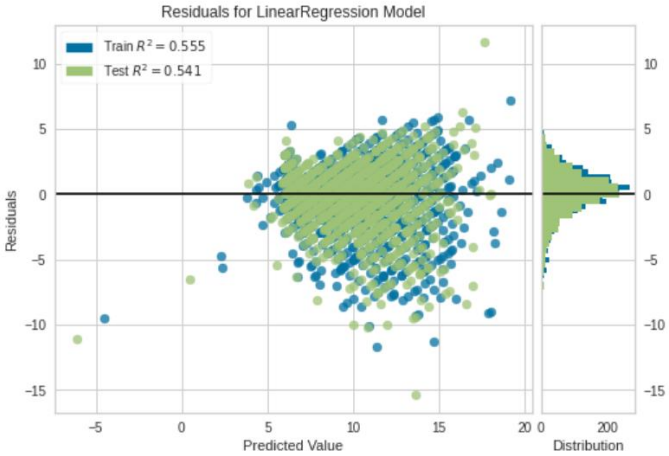

Figure 9.2: Residuals for Case 2

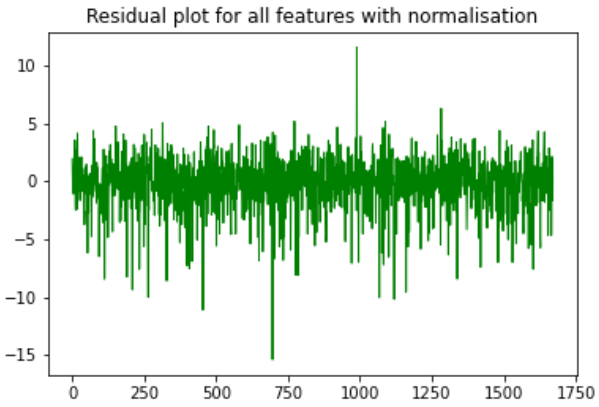

Figure 9.3: Residual plot for Case 2

We expected that the results of the prediction would improve when we normalise the features of our model. Here we have observed in Figure 9.1 that the identity line has moved closer towards the best fit line, indicating a better performance of our model, proving our assumption to be correct.

On performing a single experiment, we observed that the RMSE value is 2.17526722 and the R2 score obtained is 0.54104079.

The R2 score obtained after normalisation is higher than that observed without normalisation, proving that this is a better model.

After performing 30 experiments, we have observed that the Mean RMSE score is 2.187076 and the Std RMSE is 0.04532949. The mean R2 score is 0.5409714 and std R2 score is 0.014110732045 indicating an overall better performance than Case 1.

Case 3: Using 2 features without Normalising input data:

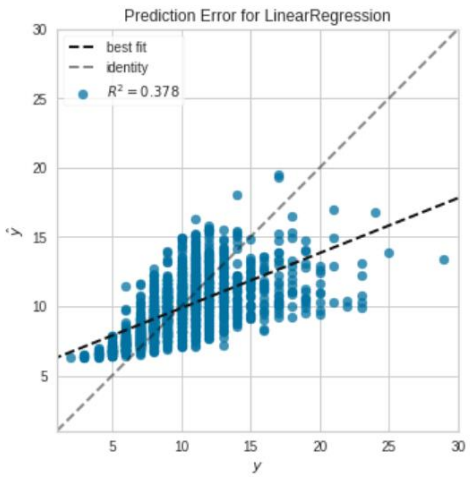

Figure 10.1: Prediction Error for Case 3

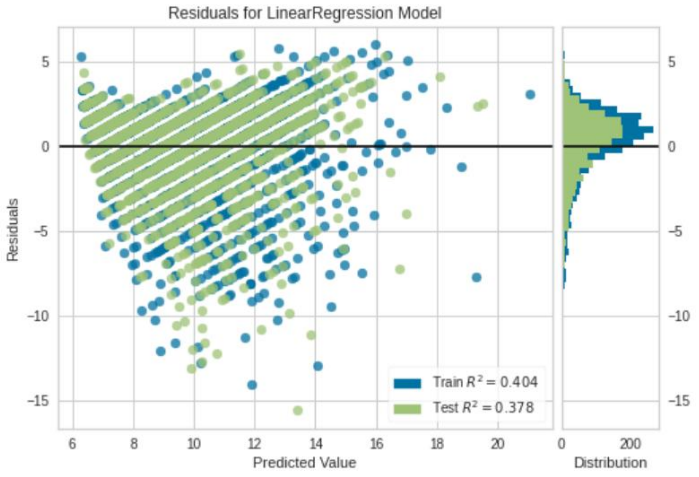

Figure 10.2: Residuals for Case 3

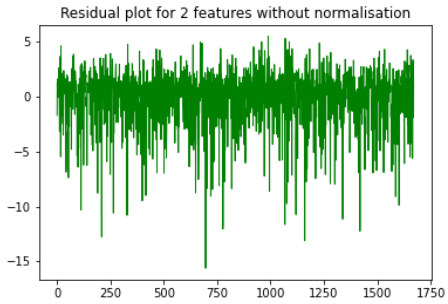

Figure 10.3: Residual plot for Case 3

Here we can observe in Figure 10.1 that the identity line is far from the best fit line, indicating poor performance of the model. Further when we calculated the metrics for the single experiment, we found that it has a high RMSE value of 2.53302 and a very low R2 score of 0.377658 which shows that this model explains only around 38% of the variation.

After conducting 30 experiments:

The metrics obtained were 2.52134 Mean RMSE and 0.05951 Std RMSE. We also approximated the R2 score as 0.390007 as Mean R2 and 0.0190617 as Std R2 score. It appears that the R2 score has improved slightly when we take the mean value from 30 experiment results.

Case 4: Using 2 features with Normalising input data:

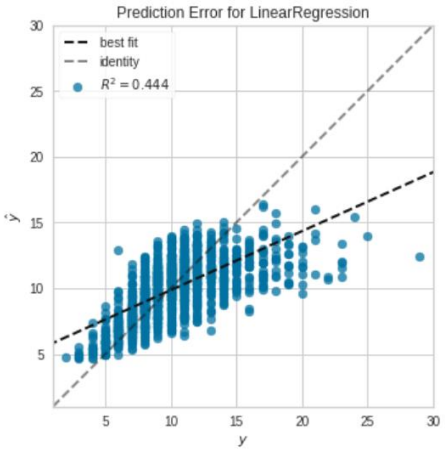

Figure 11.1: Prediction Error for Case 4

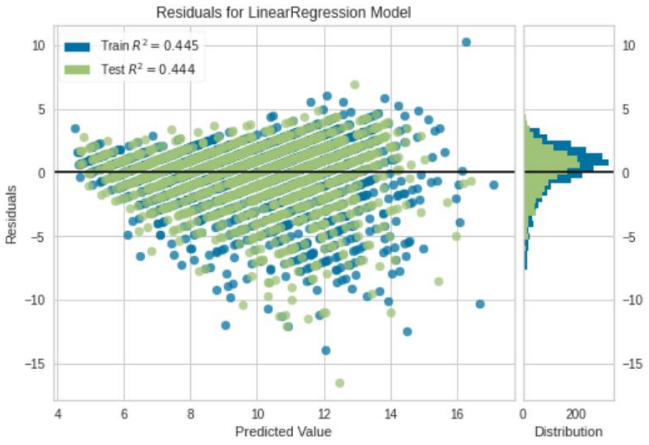

Figure 11.2: Residuals for Case 4

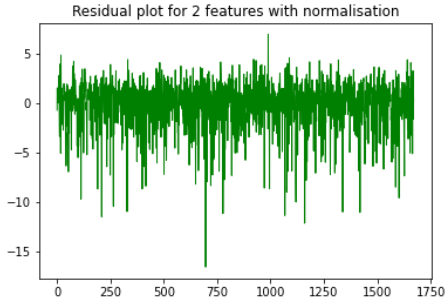

Figure 11.3: Residual plot for Case 4

Once again, we observe that the performance has slightly improved after normalising the data. In Figure 11.1, the identity line has slightly tilted towards the best fit line.

The metrics obtained were 2.39505 RMSE value and 0.443611 R2 score when a single experiment was conducted.

After performing 30 experiments:

We observe that the metrics are 2.409618 Mean RMSE value and 0.054745 Std RMSE value. We also calculated the R2 score for each experiment and obtained Mean R2 value of 0.4429107 and Std R2 as 0.015165.

Discussion of Results and Conclusion:

	Mean RMSE	Std RMSE	Mean R2 score	Std R2 score
Case 1: All features without normalisation	2.2374	0.0512	0.5193	0.0224
Case 2: All features with normalisation	2.1871	0.0453	0.5409	0.0141
Case 3: Two features without normalisation	2.5213	0.0595	0.3900	0.0190
Case 4: Two features with normalisation	2.4096	0.0547	0.4429	0.0152

Table 1: Comparison of metrics for 4 different cases

A multiple linear regression model was used to predict the Ring-Age of Abalones under two scenarios – with all input features and with two input features. For each of these scenarios we tested out the effect of normalisation of data, giving rise to 4 different cases as shown in Table 1 above.

In general, we can see that the performance in each scenario (i.e., Using all features or using two features), the model has improved in performance when we normalised the data in each case. This can be observed from the above table that Case 2 and Case 4 have lower Mean R2 scores as compared to Case 1 and Case 3 respectively. This is because normalisation helps to bring all data variable values to the same range without causing distorting differences in range of values [4]. It is often a good practice to use normalisation when we are unaware of the distribution of the data [4].

Contrary to our assumptions, our experiment proved that the model performs better when using all features rather than using the best 2 features we had selected from the correlation matrix. Since we found that our model works better with normalisation of data, we can compare Case 2 and Case 4 here. We observe that for Case 2 the mean RMSE score is 2.1871, but for Case 4 it is slightly larger i.e., 2.4096 indicating a weaker performance. We can also see that the R2 score is 0.5409 for Case 2 which is much better than the R2 score for Case 4 which is 0.4429.

Overall, the best performance was observed in Case 2, with an R2 score of 0.5409 as the model explains 54% of the variation in the target variable. The worst performance was observed in Case 3, as it produced a very low mean R2 score of around 0.39 which explains only 39% of the variation.

Though the results of this experiment show that selecting all features give better performance than using two selected features, we cannot conclude that this will always be true. The manner in which we select the input features largely depends on the data set. Further experiments need to be conducted to see if the selection of 2 different features apart from Shell weight and Diameter, will give different results.

For future work, we can conduct the same experiment by testing different combination of input features and using different methods for cleaning of the dataset. We can also aim to compare different metrics like accuracy and Mean Absolute Error (MAE).

References:

1. Exercise 1.3 Solution on Ed: https://edstem.org/au/courses/6212/lessons/13871/slides/110007
2. Exercise 1.4 Part 1 Solution on Ed: https://edstem.org/au/courses/6212/lessons/13871/slides/111793
3. Visualising Regression models: https://towardsdatascience.com/visualizing-linear-ridge-and-lasso-regression-performance-6dda7affa251
4. https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff