ZZSC5836 Assessment 2, Option I

*Submitted in partial fulfillment of ZZSC5836

1st Paul John Cronin
Department of Science
University of NSW
Sydney, Australia
p.cronin@student.unsw.edu.au or 0000-0002-4531-1481

***Abstract*—This document considers the questions raised in Assessment 2 of ZZSC5836 Machine Learning and Neural Networks. This assessment offers the choice between two different options; this documents answers "Option I: Data processing and linear regression".**

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

The purpose of this paper is to predict the age of abalone from supplied physical measurements, those being linear measurement (abalone Length, Diameter and Height), mass measurement (abalone Whole, Shicked, Viscera and Shell mass) and the categorical classification of Sex (if the abalone is Male, Female or Infant).

This paper will examine the use of linear models to achieve the goal of predicting abalone Age.

II. DATA PROCESSING

A. Clean the data

The first step was to clean the data, as the abalone comes in three categorical possibilities - 'Male', 'Female' and 'Infant'. Following instructions from the assessment, Male abalone was designated '0' and Female abalone was designated '1'. Following instruction from the assessor[1], the Infant data was classified as '-1'. There are other ways to encode this categorical data, such as with a hot-one encoding[2].

B. Exploratory data analysis

Just for the purpose of exploring the data, the total dataset was first split into three datasets based upon categorical Sex values of Male, Female and Infant, as well as an additional two categories, that being abalone younger than 10.5 years (designated 'Young') and abalone older than 10.5 years (designated 'Old'). The following table describes these data set breakdowns when looking at Age.

Table I: Exploratory Data Analysis

Statistic	*Male*	*Female*	*Infant*	*Young*	*Old*
count	1528	1307	1342	2730	1447
mean	10.71	11.13	7.89	8.14	13.31
std	3.03	3.10	2.51	1.60	2.77
min	3.00	5.00	1.00	1.00	11.00
median	10.00	10.00	8.00	8.00	12.00
max	27.00	29.00	21.00	10.00	29.00

There are statistically significant instances of each category. As the Young and Old datasets were artificially created, it was expected that Young's maximum Age would be 10, and Old's minimum Age would be 11. What was unexpected was that Infant abalone had a maximum age of 21, with more than 155 of its 1342 instances (11.54 percent) being greater than 10 years of age. Apart from these outliers, the Infant dataset seems to approximate the Young dataset. Further, both the Male and Female abalone had minimum ages of 3 and 5 years respectively - which perhaps should have categorized them as Infants. From this analysis, it was clear that Sex may not contain valuable information as the Male and Female datasets contain young members, and the Infant dataset contains older members. The dataset contains two types of explanatory variable: linear measurements (Length, Diameter and Height) as well as mass measurements (Whole, Shucked, Viscera and Shell weights). In general, the mass of any object goes as the cube of its length:

$$mass \sim length^3 \tag{1}$$

[1]https://edstem.org/au/courses/8454/discussion/846147?comment=1907849

[2]https://edstem.org/au/courses/8454/discussion/846147

C. Heatmaps

As per the assessment, a correlation heatmap was developed - see Figure 1a. A number of insights can be drawn from this heatmap:

- As would be expected, there are strong correlations between some of the linear measurements (Length and Diameter), as well as between some of the mass measurements (Whole, Shucked and Viscera).
- Somewhat unexpectedly, the Height measurements do not correlate as well with the other linear dimension measurements (Length and Diameter).
- Perhaps most importantly, as the goal is to develop a linear model to predict the age of the abalone, none of the predictors independently correlate well with Age; all are less than 0.63.

It is for these reasons, additional heatmaps were generated for Figures (1b) Male, (1c) Female, (1d) Infant, (1e) Young and (1f) Old datasets. This gave further insights into the data:

- The Male and Female abalone data was not dissimilar to the All abalone dataset, with the Male data correlating slightly better than the Female dataset.
- The Infant and Young datasets were characterized by significantly better correlations for all variables, including much stronger correlations of all variables with Age.
- The Old abalone dataset was characterized by similarly good correlation between the explanatory variables, but

(a) All abalone

(b) Male abalone

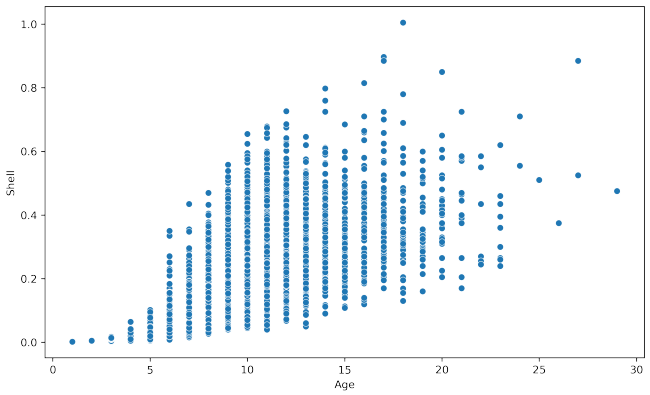
(c) Female abalone

(d) Infant abalone

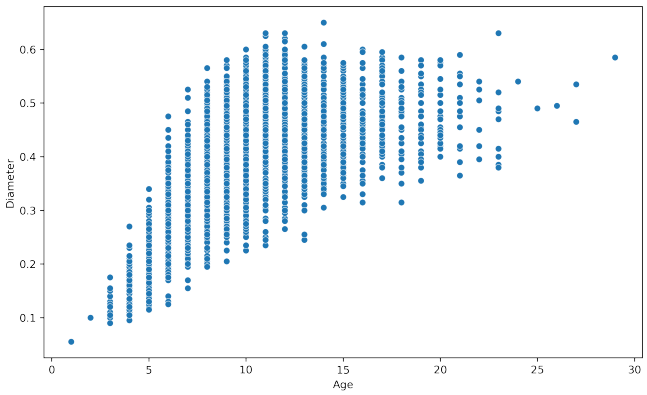
(e) Young abalone

(f) Old abalone

Figure 1: Heatmaps showing correlation between various abalone predictor variables

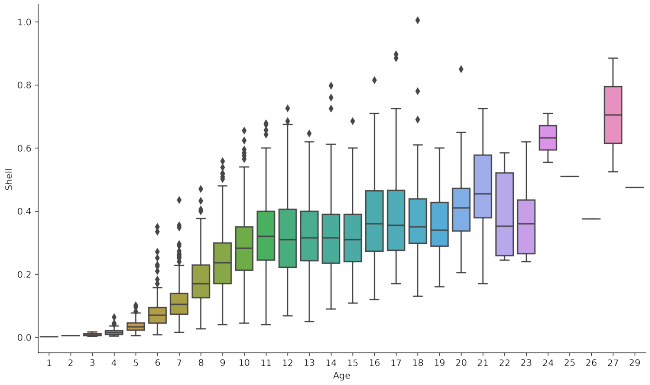


(a) Shell mass vs Age

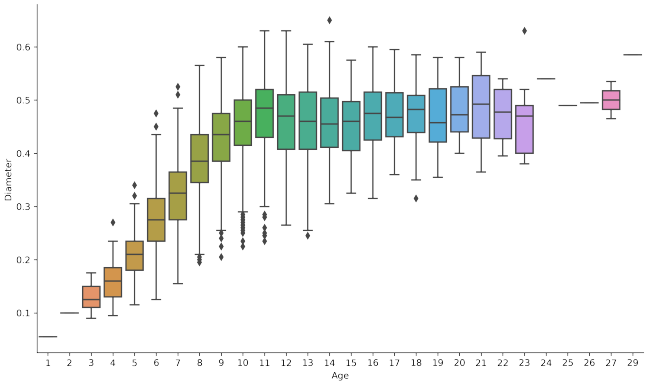


(b) shell Diameter vs Age

Figure 2: Abalone scatterplots for Shell mass and shell Diameter



(a) Shell mass vs Age



(b) shell Diameter vs Age

Figure 3: Abalone categorical plots for Shell mass and shell Diameter

there was very little correlation between the explanatory variables and the Age variable. The best correlating variable was Shell mass, with a correlation coefficient of only 0.211. The rest of the variables had almost no correlation with Age at all.

D. Correlated features

Following the assessment requirements, the two most correlated features with abalone Age were Shell mass (0.628) and shell Diameter (0.575). Two scatter plots were generated from these variables with Age - see Figures 2a and 2b.

There is little insight to be gained from this form of scatter plot, except that both figures exhibit significant heteroskedasticity.

However, much greater insights into the data can be found if box-type categorical plots are generated, instead of the simple scatter plot - see the equivalent figures 3a and 3b.

From these categorical plots, the following insights are found:

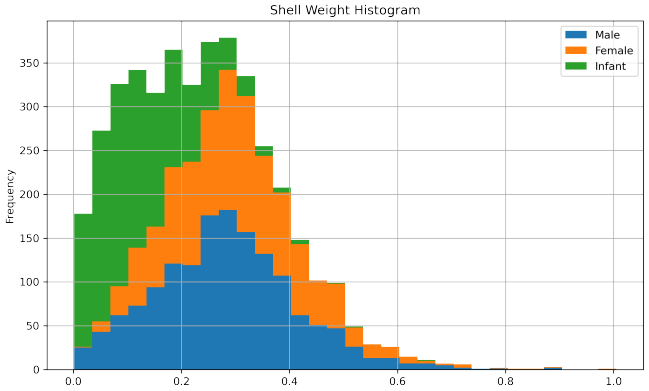
- There is a seemingly strong linear relationship between the Age and linear dimensions (Diameter, Length and Height), but only for Age less than 10.5 years. After that Age, that strong linear relationship fails, as it seems the abalone reaches maturity and does not change linear dimensions.
- There is a seemingly strong non-linear relationship between the Age and mass dimensions (Whole, Shucked, Viscera and Shell weights). As with the linear dimension, this strong relationship fails after Age greater than 10 years.
- This "maturity non-linearity" could be a significant problem for a linear model trying to predict Age, when the age of the abalone exceeds 10.5 years, or if a single linear model attempts to cover the entire range of abalone Ages.

E. Histograms

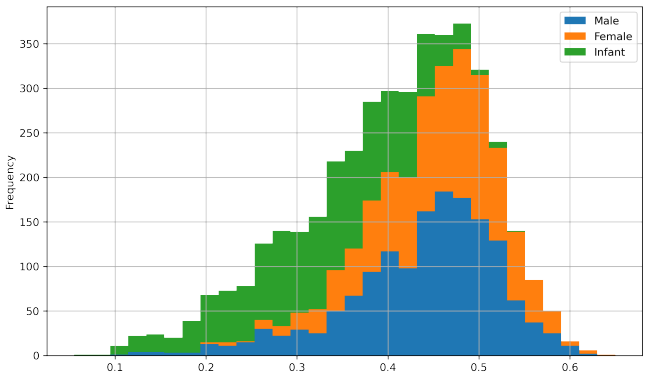
Following the assessment requirements, the two most correlated features (shell Mass and shell Diameter), as well as the ring Age were plotted as "stacked histograms"[3] where the Male, Female and Infant data is broken out. The results can be seen in figures 4a and 4b.

From these figures, the following insights were found:

[3]https://stackoverflow.com/questions/18449602/create-stacked-histogram-from-unequal-length-arrays



(a) Shell mass histogram



(b) shell Diameter histogram

Figure 4: Abalone histograms for Shell mass and shell Diameter

- That for both the Male and Female data, the shell Weight, shell Diameter and abalone Age were very similar.
- That the Infant data was usually smaller, weighed less, and not unexpectedly, was not as old.

F. Additional visualisations

In addition to the figures explicitly requested in the assessment, a number of additional visualisations were generated. While exploring the data, one of the first steps was to create histograms of the different Sex classes, in regard to Age. Initially, it was just for the three given classes (Male, Female and Infant), but then addition synthetic classes (Young and Old) were generated. It was from this initial data, that the conclusion was formed that there was little difference between the Male and Female Age distributions.

What was unexpected was that a large number of elements of Male and Female overlapped with the Infant dataset, and the Infant dataset itself extended past 20 years of age.

It was for these reasons that the synthetic data (Young and Old) were generated, to examine how Age was truly distributed.

From examining categorical plots of the various predictor variables against Age (Figure 6), a number of curious, and potentially problematic, features can be found. They include:

- that there is a non-linearity with all the predictor variables at approximately 11 years of age - the abalone simply stops changing its linear dimensions and weight. It has, effectively, come to maturity, and there is very little evidence to differentiate a 20 year old abalone with an 11 year old abalone.
- all the linear measurements of the abalone (Height, Length and Diameter) appear direcly linearly proportional to the Age of the abalone, that is, the abalone growns in its linear dimensions uniformly every year, until age 11, then changes length very little thereafter.

$$Age \sim Length \tag{2}$$

- all the mass measurements of the abalone (Shell, Shucked, Viscera and Whole mass) go as a nonlinear response:

$$Age \sim Mass^n \tag{3}$$

By considering Equations 1, 2 and 3, together a good inference is that *n=3*, which would agree with the data, such as Figure 3a.

$$Length \sim Age^{\frac{1}{3}} \tag{4}$$

III. Modelling

The section responds to the assessment questions in regard forming linear models of the data, one linear model containing all the predictor variables, including the categorical variable Sex, and another linear model containing only two of the most important predictor variables.

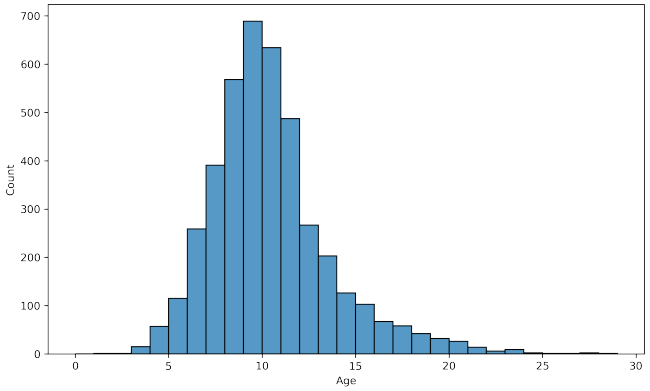
To create and validate these linear models, the data was divided into a 60/40 train/test randomised split, where each split was based upon a randomisation seed. Note: the randomized set was repeated for each modelling and normalisation approach, so that direct comparison would be possible.

Further, the data was modelled and normalised using the following schedule:

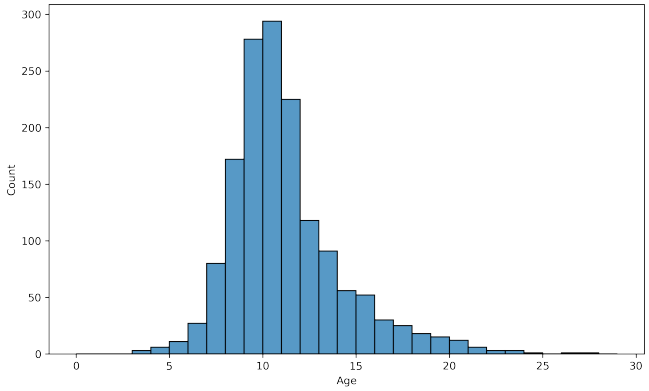
- Stochastic Gradient Descent regressor with no normalisation (SGD/No)
- Stochastic Gradient Descent regressor with the SKLearn Standard Scaler (SGD/SS)
- Stochastic Gradient Descent regressor with the SKLearn Min Max Scaler (SGD/MM)
- Stochastic Gradient Descent regressor with the SKLearn Max Abs Scaler (SGD/MA)
- SKLearn Linear Regression (Linear). For this modelling approach, it does not matter which normalisation is used, or if no normalisation is used.

In Tables II and III the results of these linear models are given, and in Table IV the linear model data is broken out by Sex category.

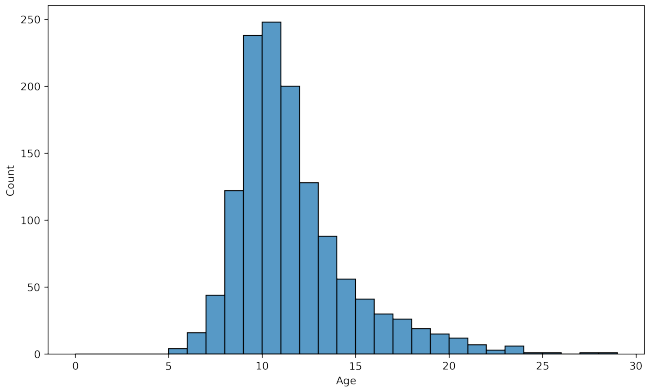
In total, 30 experiments were undertaken for each model, for each split, and for each normalisation, and the mean and



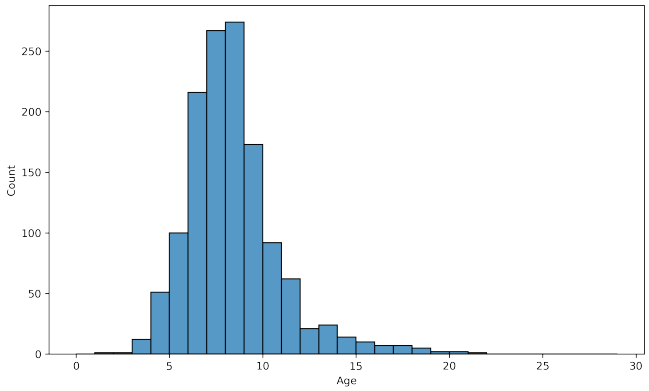
(a) All abalone



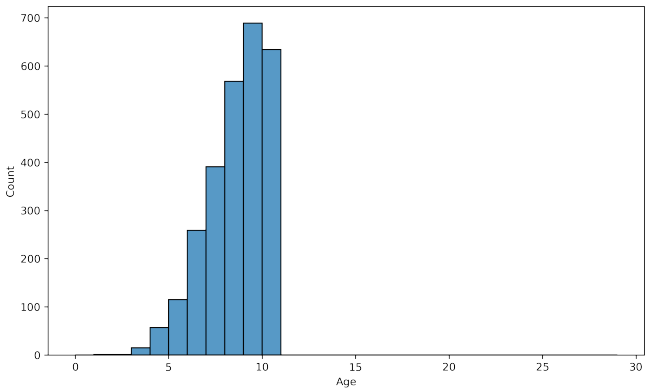
(b) Male abalone



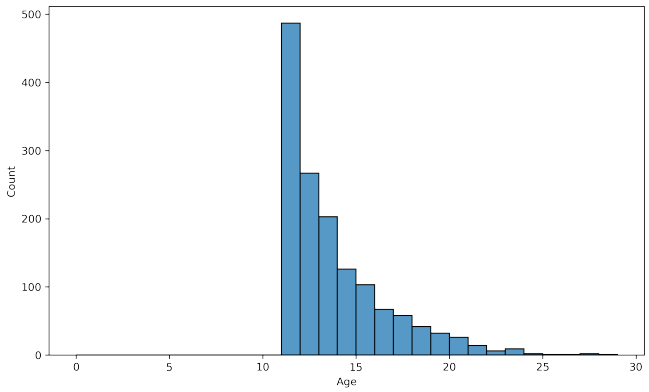
(c) Female abalone



(d) Infant abalone



(e) Young abalone



(f) Old abalone

Figure 5: Histograms of the Age distribution, broken out by Sex

standard deviation of the RMSE and R-squared score of the train and test datasets were recorded.

A. *Linear regression model using all features*

The following linear regression model was developed, using all the continuous predictor variables, as well as the categorical variable Sex to predict Age.

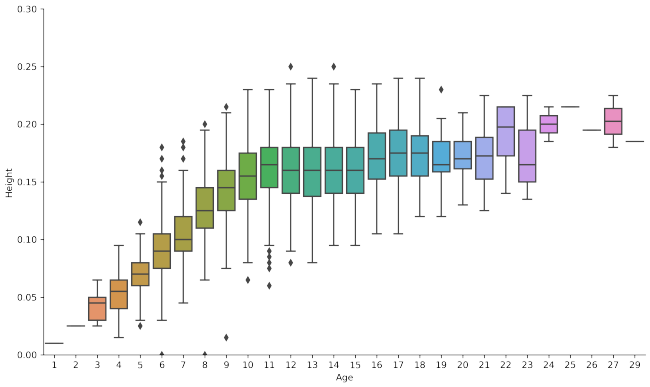
$$Age \sim Sex + L + D + H + W + Shu + V + She \quad (5)$$

where L is the abalone Length, D is the abalone Diameter, H is the abalone Height, W is the abalone Whole weight, V is the abalone Viscera weight and *She* is the abalone Shell weight.

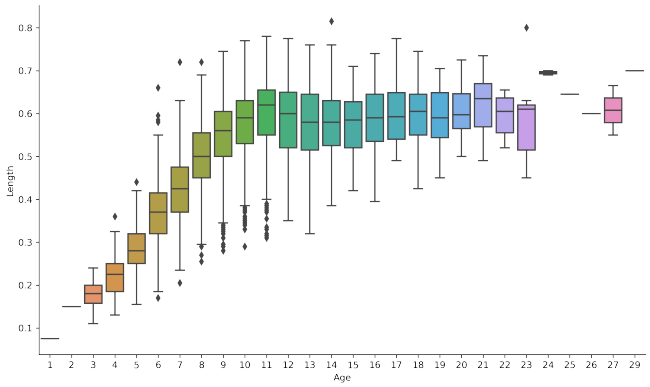
SKLearn[4] linear models were trained with all the predictor variables, and the following RMSE and R-squared values were recorded for each type of normalisation - see Table II.

Examining this table, it is clear that the best model (smallest RMSE and largest R-squared values) is the SKLearn Linear-Regression model (Linear), followed by a SGD regressor with

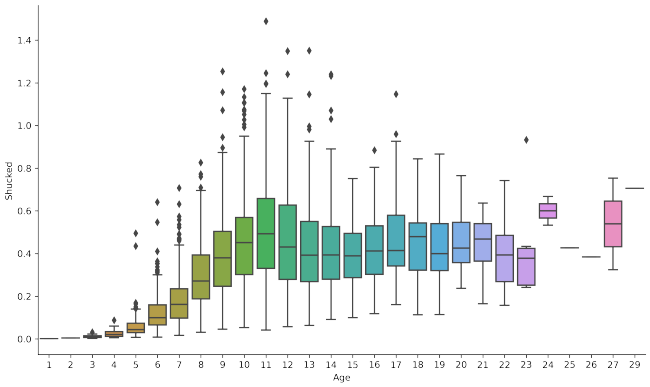
[4]https://scikit-learn.org/stable/modules/classes.htmlmodule-sklearn.preprocessing



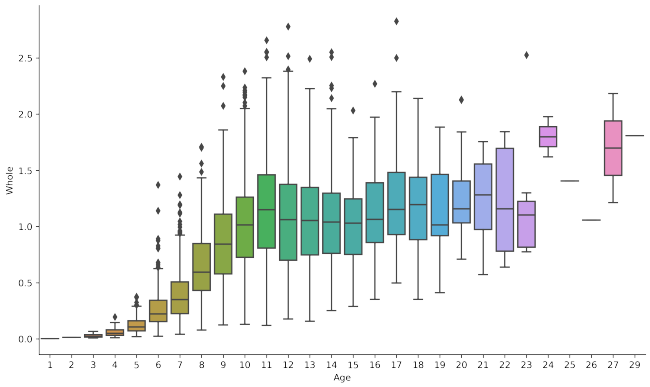
(a) Abalone height



(b) Abalone length



(c) Shucked abalone mass



(d) Whole abalone mass

Figure 6: Seaborn categorical plots of the various predictor variables with Age.

Table II: Linear Model with all features

Statistic	*SGD/No*	*SGD/SS*	*SGD/MM*	*SGD/MA*	*Linear*
μ_{RMSE}	2.288	2.253	2.347	2.345	2.237
σ_{RMSE}	0.058	0.056	0.055	0.054	0.051
μ_{R^2}	0.498	0.513	0.471	0.472	0.519
σ_{R^2}	0.015	0.023	0.022	0.022	0.022

a Standard Scaler. Both of these models can explain more than 50 percent of the variation.

However, this is still a poor overall result, and arises from the fact that there is a strong non-linearity after Age = 10 years old. One method to improve the error would be to split the data into two age groups (Young and Old), fit each model individually, and then combine the results.

B. Linear regression model with two selected features

Another linear regression model was developed, this time only using two predictor variables, those being abalone shell Diameter and the abalone Shell weight:

$$Age \sim D + She \tag{6}$$

where *D* is the abalone Diameter and *She* is the abalone Shell weight.

In the same manner as the previous SKLearn linear models were trained and normalized, so too was this model trained and normalised, but this time with only the two best predictor variables. Again, the RMSE and R-squared values were recorded for each type of modelling and normalisation - see Table III

Table III: Linear Model with two features

Statistic	*SGD/No*	*SGD/SS*	*SGD/MM*	*SGD/MA*	*Linear*
μ_{RMSE}	2.534	2.521	2.554	2.549	2.521
σ_{RMSE}	0.059	0.059	0.058	0.058	0.060
μ_{R^2}	0.384	0.390	0.374	0.377	0.390
σ_{R^2}	0.017	0.019	0.017	0.017	0.019

As would be expected when using a subset of the predictive variables (only two variables instead of the original eight), the RMSE and R-squared values are worse than before. The RMSE values are very similar for the different linear modelling and normalisation approaches, but the SKLearn LinearRegression model is still the best performing. The SGD models are somewhat poorer, but the StandardScaling normalisation is the best of them.

The R-squared values have fallen quite significantly, from more than 50 percent to below 40 percent. The LinearRegression and SGD with StandardScaling perform the best at 39

percent.

C. Linear regression model broken out by Sex

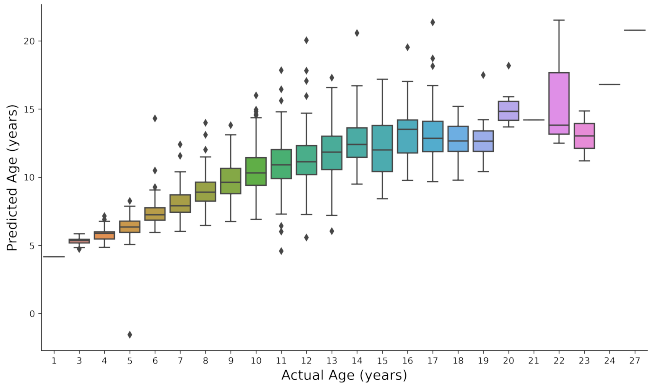
Another linear regression model was developed, this time using all the predictor variables, but broken out by the Sex category. As discussed earlier, addition Sex categories were synthetically generated, giving the following breakdown of the data:

- All abalone,
- Infant abalone,
- Male abalone,
- Female abalone,
- Young abalone (less than 10.5 years old), and
- Old abalone (more than 10.5 years old).

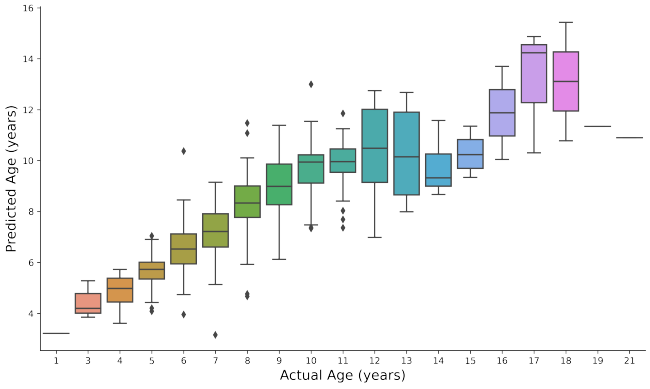
As would be expected, the All abalone RMSE and R-squared values match the LinearRegression results from Table II.

By examining Figure 7 simultaneously with Table IV, a number of insights can be gained.

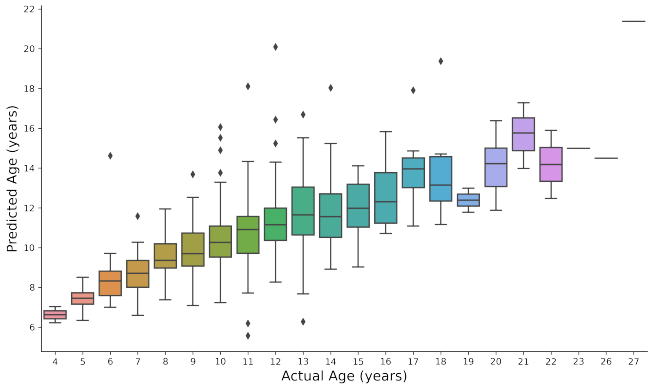
First, if we only consider the Young and Infant datasets, much smaller RMSE and much greater R-squared values are obtained - less than half the value for RMSE (1.035 vs 2.237), while the R-squared values increase to 0.583 and 0.573



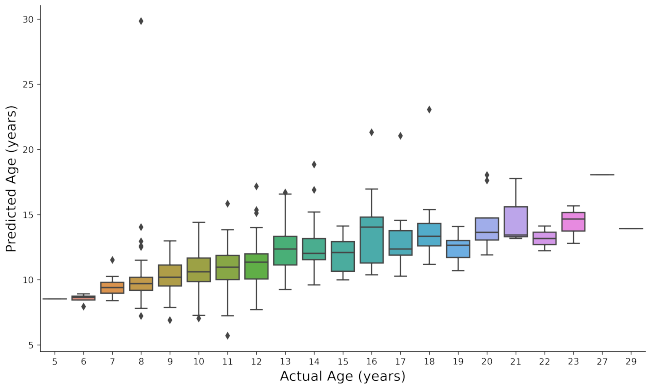
(a) All abalone sexes



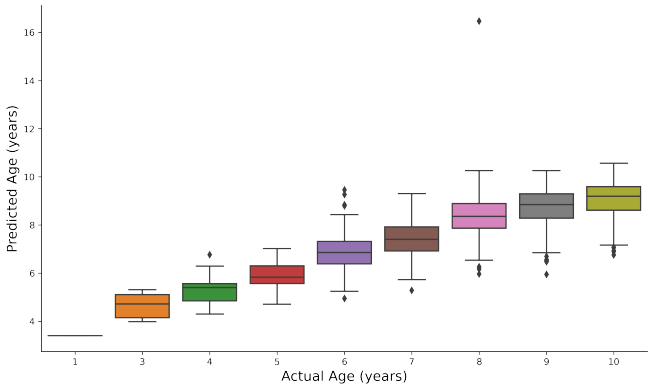
(b) Infant abalone



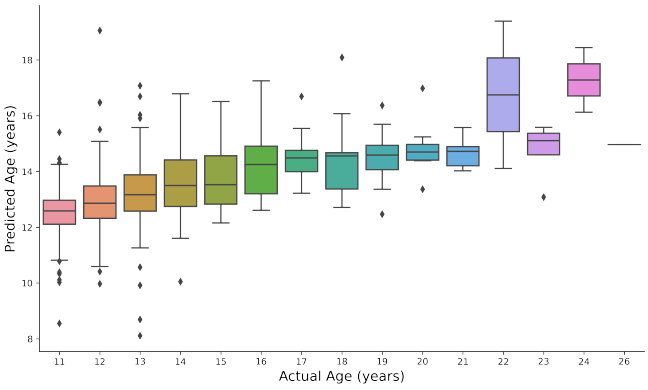
(c) Male abalone



(d) Female abalone



(e) Young abalone



(f) Old abalone

Figure 7: Seaborn categorical plots of the various predictor variables with Sex.

Table IV: Linear Model with all features by Sex

Statistic	*All*	*Infant*	*Male*	*Female*	*Young*	*Old*
μ_{RMSE}	2.237	1.638	2.303	2.552	1.035	2.401
σ_{RMSE}	0.051	0.061	0.084	0.120	0.025	0.082
μ_{R^2}	0.519	0.573	0.430	0.315	0.583	0.252
σ_{R^2}	0.022	0.030	0.030	0.045	0.029	0.030

respectively. This is understood by recalling that there is a non-linearity in the growth of abalone; above age 10, the abalone stops growing. By only considering abalone younger than that maturity cut-off, superior prediction of Age is possible.

Likewise, when considering abalone over the maturity cut-off age, the ability to predict Age from measurement declines significantly, with the R-squared value falling to 0.252. This would indicate that if an abalone was over the Age of 10 years, predictive accuracy would be very poor.

Curiously, the Male dataset performs better than the Female dataset with RMSE of 2.303 for Male and 2.552 for Female. Likewise the R-squared value for Male is better than the R-squared value for Female (0.430 and 0.315). A hint of this result was seen in the Heatmap section of this report.

IV. Conclusion

This report is in fullfillment of Assessment 2 for ZZSC5836. It covers the Exploratory Data Analysis (EDA) and linear modelling of abalone data, with the goal of accurately predicting the age of the abalone using linear and mass measurements of the abalone.

The EDA generated heatmaps of the variables, finding that linear measurements correlated well with linear measurements, that mass measurements correlated well with mass measurements, but neither correlated strongly with Age. The two most strongly correlating predictor variables with Age were Shell mass and shell Diameter.

Scatter plots and Categorical plots of the various predictor variables against Age were generated, and a serious problem was discovered; that being there was a discontinuity in the data. Around Age 10 or 11 years, abalone ceased to grow further, a strong non-linearity that would limit the accuracy of any simple linear modelling.

Histograms were also generated of the two most important predictor variables, those being Shell mass and shell Diameter, and showed expected results.

Finally modelling of the abalone data was undertaken using the SKLearn packages and various normalisations. It was found that the best results were obtained when the Linear-Regression() model was used, rather than SGDRegressor. The best normalisation technique was the StandardScalar.

When considering the data as a whole, the best RMSE was 2.237, with a R-squared value of only 0.519. This is considered a poor result.

It was possible to improve upon this result quite significantly by considering abalone under the Age of 10.5 separately from abalone over the Age of 10.5. When this is done, the RMSE reduces to 1.035 and the R-squared value increases to 0.583 for Young abalone.

Abalone older than 10.5 years have a much reduced predicted Age accuracy, with an RMSE of 2.401 and R-squared value of only 0.252.

It is the conclusion of this report that linear modelling will not give improved results over the standard slice-and-count technique, but a more sophisticated model could match that accuracy.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

Acknowledgment

The author would like to acknowledge Dr Rohitash Chandra, the course instructor for ZZSC5836. The Jupyter Notebook that accompanies this report was significantly assisted by the course work[5].

[5]https://edstem.org/au/courses/8454/lessons/20312/slides/144652