Regression Modelling in Predicting the Age of Abalone: A Comparative Analysis

Minoli R. Munasinghe
Department of mathematics and Statistics
Thompson Rivers University
Kamloops, Canada
minolimunasinghe@outlook.com

Abstract

The purpose of this study is to investigate the effectiveness of different regression models for predicting the age of abalones. It has been shown that abalone has a positive correlation with its age in terms of the economic value [1]. Therefore, precisely determining the age of abalone holds significant importance in the fisheries sector for making informed decisions regarding pricing. Many researchers have been conducted studies on predicting the age of abalone using its physical measurements such as length, diameter, height, shell weight, etc. This study examines the performance of several different statistical and machine learning models such as multiple linear regression model, Principal Component Regressor (PCR), Lasso regression model, Ridge regression model, Support Vector Regressor (SVR), Random Forest Regressor (RFR), etc. on abalone age prediction using the Abalone dataset available in UCI machine learning repository. The performances of the models are evaluated using computational efficiency and various statistical measures such as Root Mean Squared Error (RMSE) and R-squared values. Furthermore, the study explores the impact of feature engineering techniques, such as scaling and data transformation, on the performance of the models. The results suggest that the Random Forest Regressor with feature engineering techniques such as scaling and log transformation outperforms the other models considering the RMSE, R-squared values and computational time.

1. Introduction

Abalone is a type of marine mollusk that belongs to the family Haliotidae. They are commonly referred to as ear shells, mutton fish, or mutton shells [2]. Abalones are highly valued for their large, fleshy body, which has made them a popular food item across many cultures worldwide. The interior of the shell is composed of a mother-of-pearl layer that displays a variety of colors, and this attractive feature has led to the use of abalone shells as jewelry items or decorative objects by humans [2]. Therefore, predicting the age of abalones is a crucial factor for decision-making in the fishing industry, owing to their high economic value globally. The traditional way of identifying the age of an abalone is by counting the number of rings on the shell. However, since this method is time consuming, many researchers in the world has been conducted research on predicting the age of the abalone using machine learning algorithms. There are numerous studies performed on abalone age prediction using classification and regression approaches.

As an example, a study has been conducted to determine the econometric ways to estimate the age and price of abalone. This study has targeted mainly on two types of models; least square estimation model and ordered probit model. According to the Least Squares Estimate method, their proposed model has omitted most of the predictor variables and it only has log of weight and height variables as predictors. The Ordered Probit Model is obtained by categorizing the number of rings into 3 classes and the variables length, diameter, height, and log of weight has been considered in fitting the model.

A research study done by a group of researchers have been identified that using typical machine learning models on predicting the age of abalone is not giving better results and therefore they have developed an ensemble model which works with both lazy and eager machine learning algorithms. The results have been revealed that the new model is performing well compared to traditional machine learning models and the performance of the model has been increased to 90.44% of accuracy [4].

Another study has been conducted by Jabeen and Ahamed on predicting the age of the abalones using Artificial Neural Network (ANN). They have developed a feed forward multi-layer perceptron network with Levenberg-Marquardt Backpropagation algorithm. According to their findings, the error rate corresponding to the target value has been reduced when there is an increase in the number of hidden layers which resulted in high performance of the model [5].

Similar to my research study on predicting the age of abalone using regression modelling, a study has been performed by Kunj Mehta to predict the abalone age using regression. Various machine learning algorithms such as Ordinary Least Squares, (RANdom SAmple Consensus) RANSAC, Huber, and Ridge regularization has been incorporated in this study to predict the abalone age. According to the findings, the RANSAC regression model has outperformed all the other models with a Mean Absolute Error (MAE) of 1.33. In addtion, the Huber regressor has become the second well-performed model with an MAE of 1.39. However, it is identified that the ridge regularized regression model didn't perform well and cross validation, SMOTE techniques didn't improve the performance [6].

Similar to the previous studies with Neural Networks, another study has been performed to predict the age of abalone using a multi-layer neural network. The proposed model in this study has obtained a testing accuracy of 92.22%. The model is consisted of 5 layers with different nodes and hyperparameter tuning has performed to obtain the optimum learning rate of 0.06. According to the findings, most important predictors in the model are shell weight, shucked weight, Whole weight, and Viscera weight.

A similar work has been carried out in a GitHub repository and it further proves my findings about predicting the age of the abalones [8]. According to the previous studies, most of the researchers has been more focused on developing machine learning models. However, majority of the research works has developed Artificial Neural Networks to predict the age of the abalone. This research study is focused on building several different machine learning models such as multiple regression model, Principal Component Regressor, Ridge Regularization model, Lasso Regularization model, Support Vector Regressor Model, Random Forest Regressor, Neural Network, and Bayesian Regression models are fitted, and a comparative analysis is performed to determine the best predictive model with less time consumption. Moreover, cross validation and hyperparameter tuning is performed to generalize the model well on the data and finding the optimum parameters. My findings on this study could help to get a good understanding about the performance of the models on the abalone data and making accurate predictions in decision making.

2. Data

The abalone dataset is collected from the UCI machine learning repository and the URL to the dataset is provided in references section. It has 4177 observations with 9 physical measurements such as length, height, diameter, weight of the shell, weight of the whole abalone, number of rings, etc. According to the data source, the data comes from an original study and the original owners have removed the missing values [3].

2.1 Variables of the Data

The physical characteristics of the dataset and its description is provided below in Table 01[3].

Table 1:The Description of the Data

Physical Measurement	Data Type	Measuring Unit	Description
Sex	Categorical	-	Belong to Male, Female, or Infant
Length	Continuous	mm	Longest shell measurement
Diameter	Continuous	mm	Perpendicular to length
Height	Continuous	mm	Height with fleshy body
Whole Weight	Continuous	grams	Weight of whole abalone
Shucked Weight	Continuous	grams	Weight of the fleshy body or meat
Viscera Weight	Continuous	grams	Gut weight after bleeding
Shell Weight	Continuous	grams	Weight of the shell after dried
Rings	Integer	-	Number of rings + 1.5 years = Age

2.2 Primary Analysis

The variable “Rings” is considered as the response variable and all the other variables can be considered as predictor variables. The distribution of the response variable, rings is obtained from a histogram as shown in figure 01. According to the results, the number of rings ranges from 1 to 29, however, majority of the abalones have 5 to 15 rings. The number of rings follows a slightly positively skewed distribution.

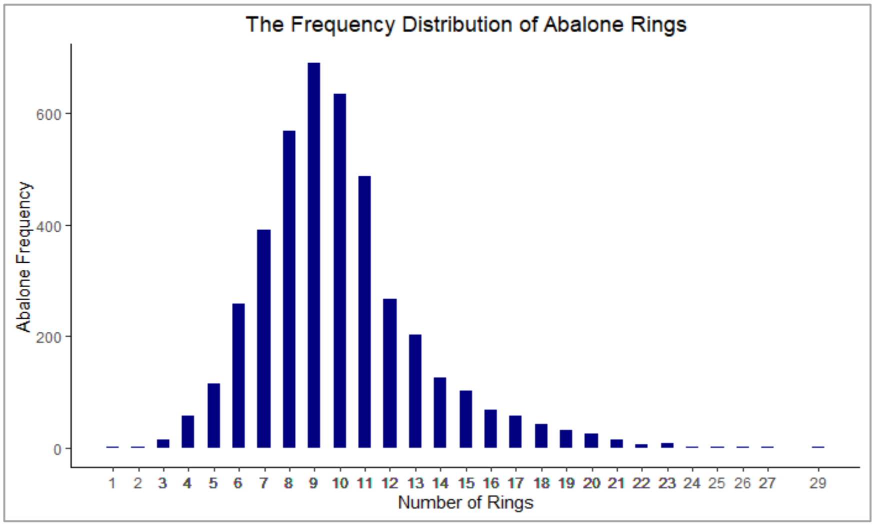

Figure 01: The Distribution of Number of Rings

The distribution of all the predictor variables is observed using a boxplot diagram as in figure 02. Accordingly, the predictor variables have different means and variations. For instance, the variable of whole weight has a greater average and variability in comparison to height, which has relatively lower values for both measures. Moreover, all the variables have outlier data which needed to be handled before the model fitting. Since a relatively small percentage of the observations in the dataset are identified as outliers, specifically 164, they have been removed from the dataset to ensure the consistency and accuracy of the data.

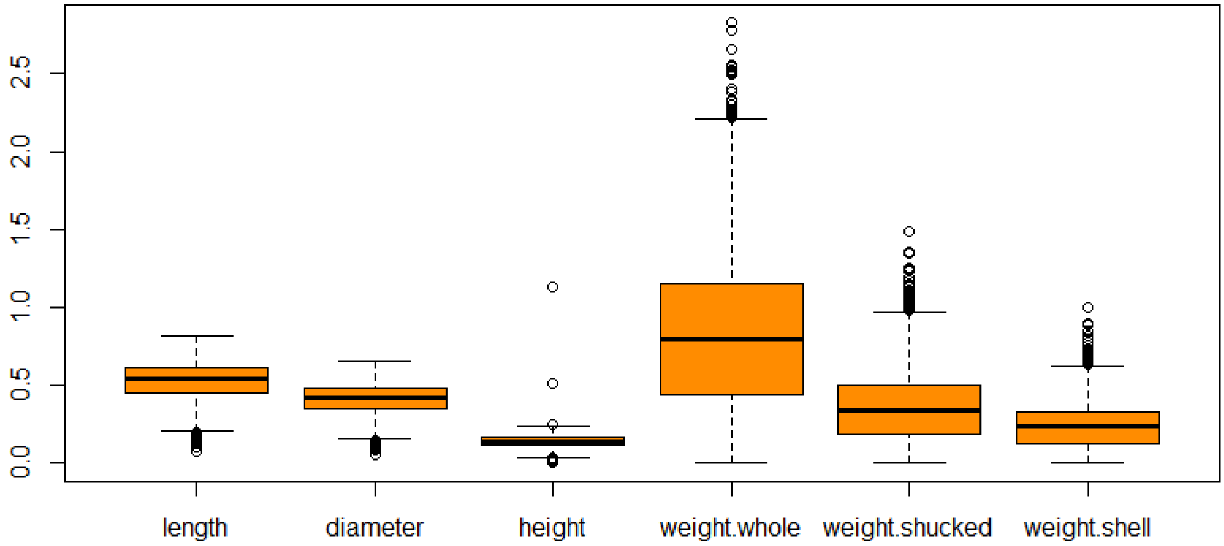

Figure 2: The boxplots of Predictor Variables

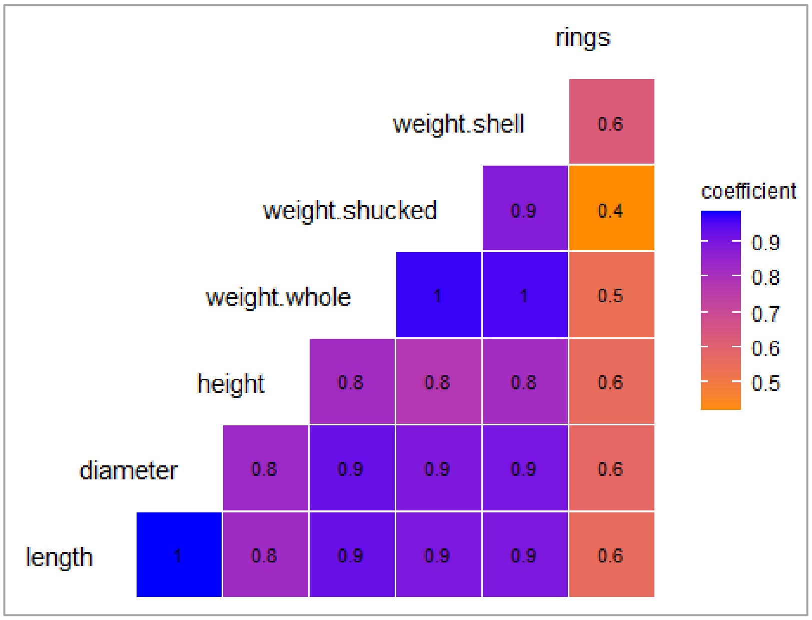

Figure 3: Correlation Plot of Variables

The correlation among all the variables is obtained using a correlation plot as shown in figure 03. According to the correation plot, there is a high multicolinearity between the predictor variables. The variables of length and diameter, whole weight and shucked weight, and whole weight and shell weight exhibit a very high correlation with each other, with a correlation coefficient of 1. Therefore, it is important to handle multicolinearity while fitting the model as it can cause inaccurate and uninterpretable results.

3. Method

3.1 Data Transformation

A multiple linear regression model is fitted using the available form of the data and observed that the assumptions are slightly violated in the model. Therefore, the response variable is transformed into its log scale, which reduce the range of the data. A multiple linear regression model is fitted again with the log(rings) as response and observed a comparatively low RMSE and high R squared value with valid assumptions. Therefore, log transformation of the response variable is used to fit all the machine learning models in the study. Moreover, the categorical variable "Sex" is transformed into a numeric variable, before fitting the regression models.

3.2 Data Standardization & Splitting

The feature engineering technique, standard scaling $((x - mean(x))/sd(x))$ is applied to all data to ensure that all the variables have similar range of data values which can improve the performance of the models.

The dataset is splitted into two separate sets, training set and testing set such that the training set contains 80% of the data whereas the testing set contains 20% of the data. The training set of data is used to train the model, while the testing set of the data is used to predict the data using the trained model.

3.3 Model Selection

Table 2 provides a list of the regression models that were chosen for the comparative analysis, along with an explanation of the reasoning behind their selection and techniques used in improving the performance of the model.

Table 2: List of Models

Model Number	Model	Reasoning & Techniques
Model 1	Multiple Linear Regression Model	▪ The basic regression model is fitted to determine the relationship between response and predictor variables and to identify the significant predictors in the model.
Model 2	Principal Component Regressor (PCR)	▪ PCR is fitted to handle the existing multicolinearity between the predictors using dimension reduction. ▪ 10-fold cross validation with hyper parameter tuning is performed to find the optimum number of components in the model.
Model 3	Ridge Regressor	▪ Ridge regressor is used for variable selection and to handle the existing multicolinearity between the predictors. ▪ 10-fold cross validation with hyper parameter tuning is performed to find the optimum lambda (regularization parameter) that reduce the

		variance of the residuals and a new model is fitted using the best lambda.
Model 4	Lasso Regressor	▪ Lasso Regressor is used to reduce the complexity of the model and to handle the exisiting multicolinearity between the predictors and select the most important predictors. ▪ 10-fold cross validation with hyper parameter tuning is performed to find the optimum lambda that maximize model performance.
Model 5	Support Vector Regressor	▪ Support Vector Regressor is used as it is capable of capturing the non-linear relationships between response and the predictor variables. ▪ 10-fold cross validation with hyper parameter tuning is performed to find the C hyperparameter which is a trade off between maximizing the margin and minimizing misclassification error. ▪ The Radial Basis Function is used as the kernel and therefore sigma, the common feature of RBF kernal is used as another hyperparameter to be tuned.
Model 6	Decision Tree Regressor (DTR)	▪ Decision tree regressor is fitted to deal with interactions between the predictors and minimize the residual variance. ▪ Grid search using 10-fold cross validation is performed to tune the complexity parameter (C_p) of the model.
Model 7	Random Forest Regressor (RFR)	▪ Deal with multicolinearity in data and prevent overfitting while minimizing the residual variance. ▪ 10-fold cross validation is performed to improve the performance of the model and selected number of predictors is tuned.
Model 8	Neural Network	▪ Neural Network is used as it is capable of determining complex relationships between the predictors and response in noisy data . ▪ Grid Serach using 10-fold cross validation is performed to obtain the hyperparameters; optimum number of hidden layers (size) and decay which is the penalty term on the neural network
Model 9	Bayessian Regression Model	▪ A prior distribution is used to measure the uncertainty of the data and update the belief accordingly.

The GitHub URL is attached for further reference of the analysis. https://github.com/minolirm/Abalone-Age-Prediction-using-Regression-Comparative-Analysis

4. Results

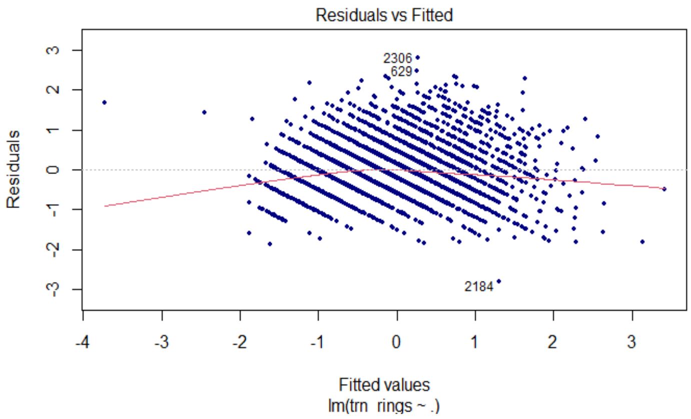

Figure 4: Residual vs Fitted Plot of the MLR model

The normal Q-Q plot and the residual vs fitted plot is obtained from the multiple linear regression model fitted using the log transformation of the response variable. The red horizontal line in figure 04 is approximately close to zero, and therefore the assumptions of equal variance and normality of the residuals is satisfied. The normal Q-Q plot in figure 5 shows that the data follows a normal distribution. Therefore, the assumptions of the multiple linear regression model are satisfied.

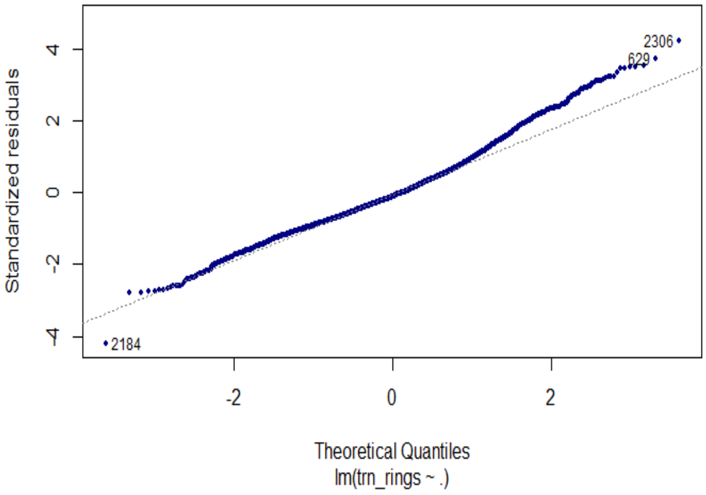

Figure 5: Normal Q-Q plot of the MLR model

The performance of each of the model is evaluated using Root mean Squared Error (RMSE) and R squared values. The results of each of the models are represented in the figure 6 and 7 respectively to perform a better comparison among the models.

According to the results, Random Forest Regressor has the highest R squared value of 0.5721 and lowest RMSE value of 0.6295. However, Support Vector Regression model has a performance similar to Random Forest Regressor with a R squared value of 0.5715 and RMSE of 0.6301. The Table 3 represents the summary results of the fitted models with optimum hyperparameter values, and the execution time of each model on training. Considering the execution time of the best two models, we can suggest that the random forest regression

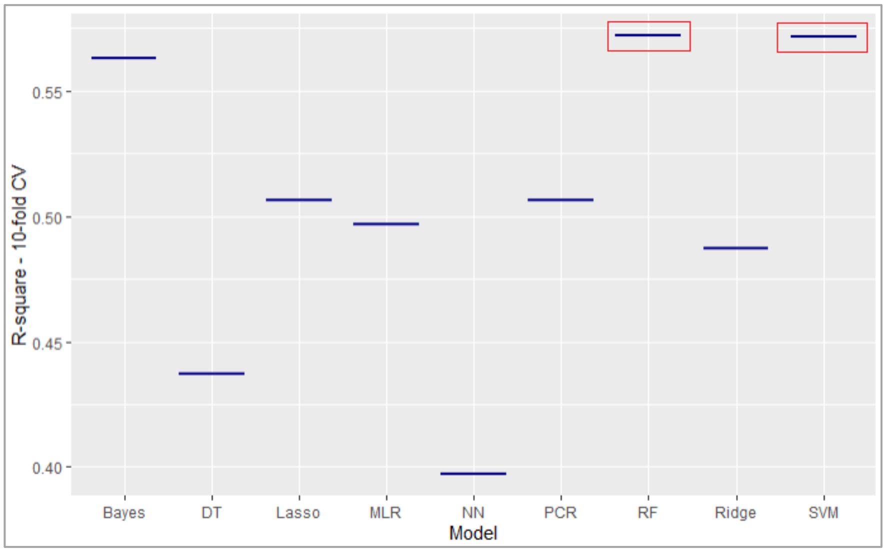

Figure 6: Boxplot Distribution for R square value of the models

model performs better than the support vector regression model with less computational cost.

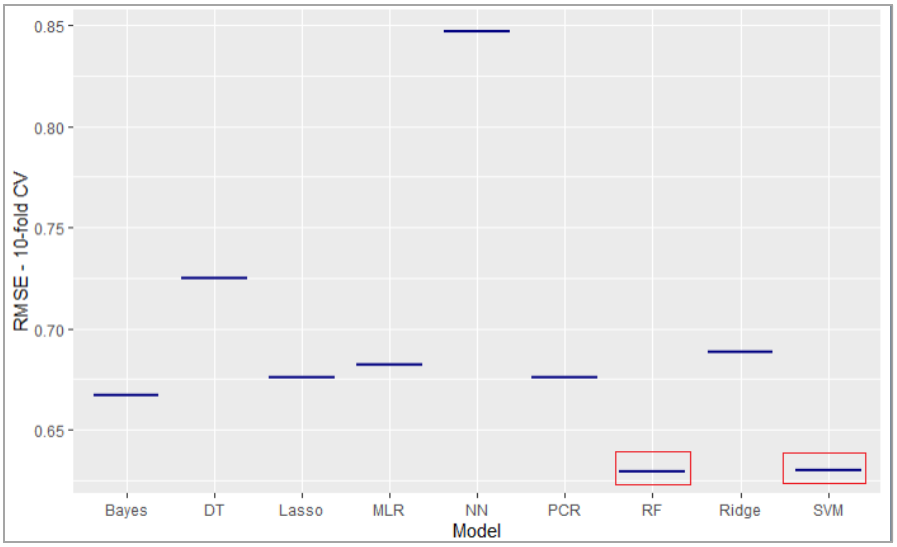

Figure 7: Boxplot distribution for RMSE of the models

Table 3: The tuned parameters and execution times of each model

Model	Optimum Hyperparameter Values	Execution time
Multiple Linear Regression model	-	0.0488 sec
Principal Component Regressor	No: of components: 8	0.6637 sec
Ridge Regressor	Optimal Lambda (λ): 0.06	2.3824 sec
Lasso Regressor	Optimal Lambda (λ): 0.00029	0.4996 sec
Support Vector Regressor	C:1 Sigma: 0.1	29.9273 min
Decision Tree Regressor	Complexity (C_p): 0.01	4.4928 sec
Random Forest Regressor	mtry (selected predictors): 2	18.3978 min
Neural Network	Number of hidden units (size):6 Weight Decay (decay): 0.1	33.3820 sec
Bayesian Regressor	-	13.1129 sec

5. Discussion

The regression modelling on predicting the age of abalones can be further expanded to fit Artificial Neural Networks with feed forward propagation and back propagation methods. Moreover, this research study would further improve by using the data augmentation techniques such as SMOTE to generate synthetic data where the groups that have a smaller number of records. In addition, combination of regularization methods with other regression models would provide better predictive results.

6. References

[1] Hossain, Mobarak, M., & Md Niaz Murshed. (2019, January 3). *Econometric Ways to Estimate the Age and Price of Abalone*. https://mpra.ub.uni-muenchen.de/91210/1/MPRA_paper_91210.pdf

[2] Wikipedia contributors. (2023). Abalone. Wikipedia. https://en.wikipedia.org/wiki/Abalone

[3] UCI Machine Learning Repository: Abalone Data Set. (n.d.). https://archive.ics.uci.edu/ml/datasets/Abalone

[4] S. Kaur, S. Chaudhary, A. Thakur, R. Bajaj, A. Gupta and A. Majotra, *"Abalone Age Prediction using Optimized Ensembel Model,"* 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 1023-1027, doi: 10.1109/SMART55829.2022.10047096.

[5] Jabeen, K., & Ahamed, K. (2016, September). *Abalone Age Prediction using Artificial Neural Network*.

[6] Mehta, K. (September 2019). *Abalone Age Prediction Problem: A Review*. International Journal of Computer Applications

[7] Mohammed, G., Shbikah, J., & Al-Zamili, M. (2021, April). *Age of Abalone Prediction from Physical Measurements Using ANN*.

[8] Nishitpatel. (n.d.). *GitHub - nishitpatel01/predicting-age-of-abalone-using-regression: Predicting the age of abalone using multiple regression in R*.
GitHub. https://github.com/nishitpatel01/predicting-age-of-abalone-using-regression

.