Predicting the Age of Abalones Based on Physical Measurements

1. Introduction to Problem and Data

1-1. Problem Statement

The ability to predict the age of abalone using physical measurements, such as sex, length, diameter, height, and weight, is an essential tool in marine biology and fisheries management. Determining the age of abalone is a time-consuming process that typically involves multiple steps, including cutting the shell, coloring it, and counting the rings (divided by 2.5) under a microscope. For my final project, I aim to develop a predictive model that can accurately estimate abalone rings based on the physical measurements of abalone. This method would provide a more efficient and convenient alternative to traditional techniques while also preventing damage to the abalone.

Abalone are precious marine resources that hold significant value to human society. They carry good food and economic values and are also significant in certain human cultures that rely on fishing for a living. The ages of abalone (calculated by the rings on their bodies) provide vital information for scientists to evaluate their growth, living conditions, reproductive patterns and population dynamics, and the health of their habitats. This model could also be applied to broader environmental issues, such as understanding how human activities could impact the growth and aging of abalones. Furthermore, these insights can guide governments in formulating policies to ensure the long-term health and stability of abalone populations and their habitats. The insights also provide scientific evidence for the sustainable development of the fishing industry. Additionally, the model can help optimize the abalone cultivation industry, increasing both the quantity and quality of abalone to meet the growing market demand. In conclusion, studying the age of abalone is not only a crucial step in advancing scientific knowledge but also holds significant implications for social economics and ecosystem balance.

1-2. Dataset Description

The data for this project is from the UC Irvine Machine Learning Repository in CSV format, providing detailed physical measurements about abalones. The data may contain null values, which require cleaning and preprocessing. Due to the complexity of biological factors influencing abalone growth, challenges may arise in constructing an accurate regression model. However, I believe that the physical measurements and categorical attributes will provide valuable insights into estimating the age of abalones with a reasonable degree of accuracy using suitable models.

This dataset contains comprehensive information about abalones and their physical and categorical attributes, including sex, length, diameter, height, weights (whole abalone/meat/gut/ dried shell), and the number of shell rings (used to determine age). A notable issue is that certain groups of variables in this dataset, such as 1. length, diameter, height, and 2. weights, seem to be evaluated from similar aspects, which implies that multicollinearity might appear and linear regression might not be applicable. It consists of 4,177 rows, each representing an individual abalone, and 8 columns on the features needed to develop and train the predictive model. These features will be profound in analyzing abalone growth patterns and estimating their age effectively.

1-3. Data Processing and Preliminary Examination

The revised dataset I will work with contains demographic information of 4,177 abalones. Our target variable, the rings on the bodies of abalones, which can be used as an indicator of abalone age, ranges from 1 to 29. These abalones can be further categorized into three different sexes: male, female, and infant (as infant abalones do not have reproductive abilities and may exhibit differences in growth rates, size, and weight to adult ones due to biological variations). The dataset also includes various physical measurements of abalones, including length, diameter, height, and weights corresponding to different parts (whole, shucked meat, viscera, and shell), and their ranges are summarized below:

Variable	*Definition*	*Range*
Length	Longest shell measurement	0.075 to 0.815
Diameter	Perpendicular to length	0.055 to 0.65
Height	With meat in shell	0.0 (minimal) to 1.13
Whole Weight	Entire abalone	0.002 to 2.8255
Shucked Weight	Abalone meat	0.001 to 1.488
Viscera Weight	Gut after bleeding	0.0005 to 0.76
Shell Weight	Shell after drying	0.0015 to 1.005

2. Exploratory Data Analysis

2-1. Descriptive Statistics

The distribution of abalone rings in this dataset is roughly normal, slightly skewed to the right, with a mean of roughly 10 and a median of 9. Meanwhile, the boxplot highlights the presence of outliers, particularly on the higher end (ring counts above ~20). These extreme values represent a small subset of abalones with an extremely high number of bodily rings, representing high age.

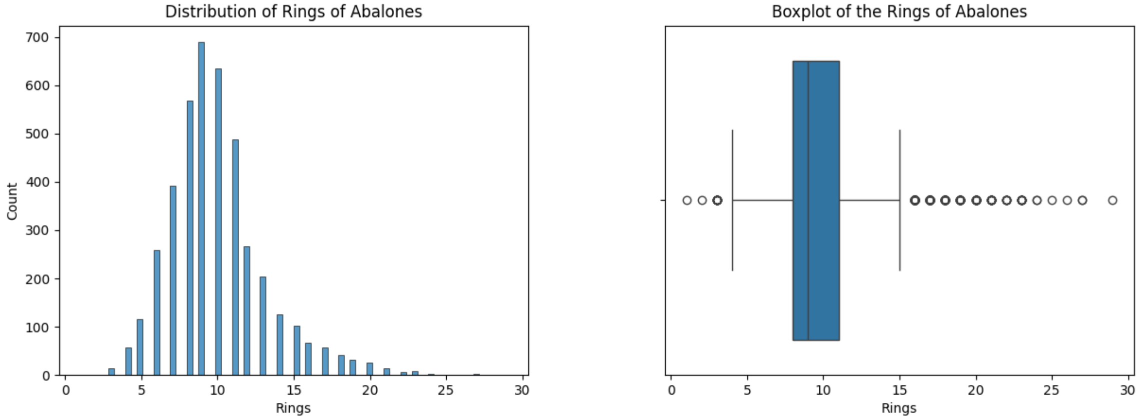


Fig. 2-1-A: Histogram for the Distribution of Abalone Rings (Ages)
Fig. 2-1-B: Boxplot for the Distribution of Abalone Rings (Ages)

2-2. Initial Visualizations

When grouping abalone rings by sex, we can discover that the average rings for male and female abalones are both above 10, with females slightly higher than males. Additionally, the average rings on infants are around 8, corresponding to around 3 years old, using the standard formula to estimate age (divided by 2.5), which is the sexually mature age of abalones.

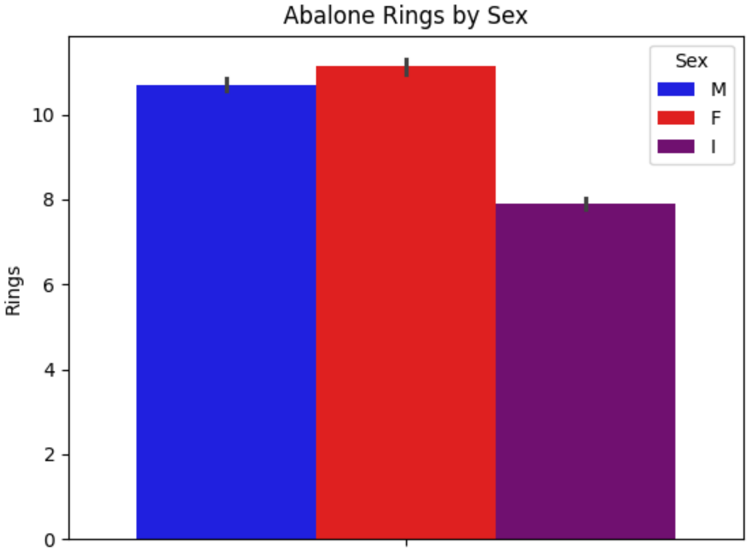


Fig. 2-2-A: Abalone Rings Grouped by Sex

The scatter plot on the next page showcases the relationship between abalone length and the count of rings on their bodies. While there is a clear positive correlation, with longer abalones tending to have more rings, the data points are relatively dispersed, particularly at higher lengths. Most abalones are clustered in the length range of 0.2 to 0.6, with corresponding ring counts primarily between 5 and 15.

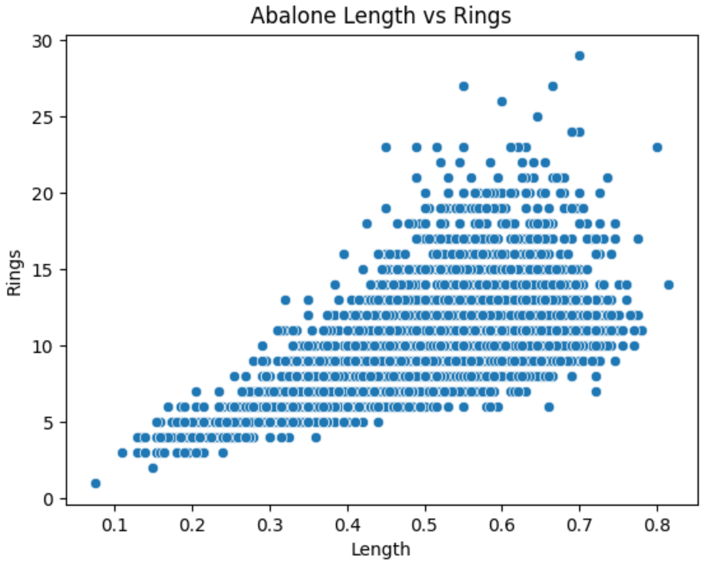


Fig. 2-2-B: Abalone Rings by Length

Concerning the multicollinearity of length, diameter, and height, the regression plot below highlights a linear relationship between the three variables. This indicates that multicollinearity does exist among them, and linear regression models would not be applicable to this scenario.

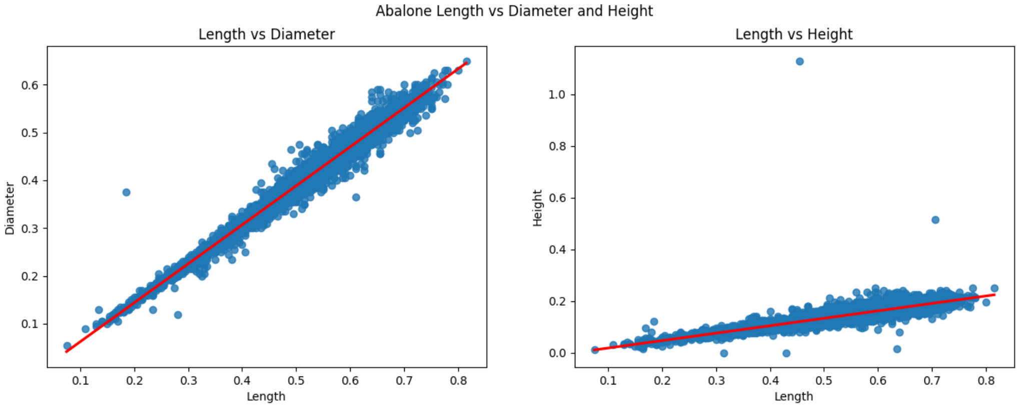


Fig. 2-2-C: Abalone Length vs. Diameter vs. Height

The scatterplot on the next page compares the whole weight of abalones to the number of rings. There is a moderate positive relationship, as abalones with greater weight generally have a higher number of rings. However, abalones with lower weights exhibit a wide range of ring counts, while abalones with higher weights (above 2.0) tend to cluster around ring counts between 10 and 20. The variability in ring counts increases with whole weight. More importantly, the scatterplot of whole weight vs rings follows a slightly different distribution to the length vs. rings scenario, which graphically implies that the intrinsic difference in the matter of these two sets of variables and that multicollinearity does not exist between the length and weight groups.

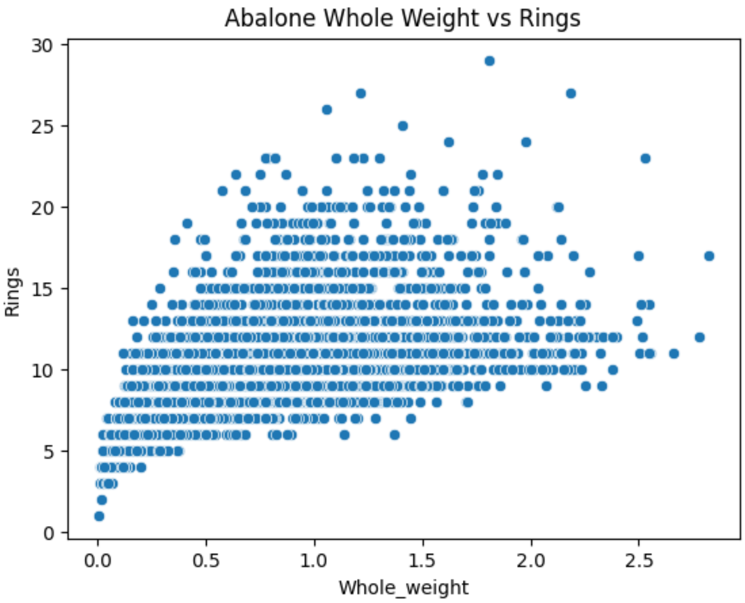


Fig. 2-2-D: Abalone Rings by Weight

Similarly, the regression plot of whole weight vs. shucked weight vs. viscera weight vs. shell weight highlights the multicollinearity among these variables, emphasizing again the inapplicability of linear regression models.

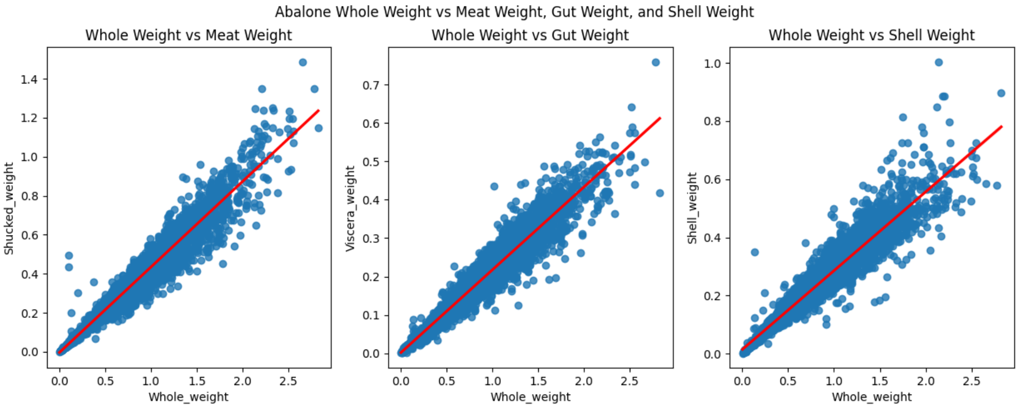


Fig. 2-2-E: Whole vs. Meat vs. Gut vs. Shell Weights

3. Modeling and Predictions

3-1. Modeling Overview

To predict abalone rings, which can be further used to infer abalone ages, I decided to use different regression models and see which one performs the best in predicting the rings and accounting for the variation in my data. Due to the multicollinearity issues existing in the dataset, I chose not to use Linear Regression Models. I decided to utilize an 80-20 train-test split for each of these models, training my model on 80% of the data and then testing it on the remaining 20%.

3-2. Baseline Model

I evaluated the success of each of my models by comparing its performance metrics, such as the model's mean squared error and R^2, against this baseline's mean squared error and R^2 (which equals 0.0 by definition). To get my baseline value, I simply took the mean abalone rings of my dataset. The Mean Squared Error for the baseline model is **10.392777255475611**. For model comparisons, we would look for if the MSE of other models is below this value, and the lower the MSE is, the better. For R^2, if the value is greater than 0, then a greater proximity to 1 will be ideal.

3-3. K-Nearest Neighbors (KNN) Regression Model

I chose to try the K-Nearest Neighbors regression evaluation metric next because KNN makes predictions based on the similarity of instances in the feature space. If abalone ages were influenced by local patterns or clusters of similar physical measurements, such as length, weight, or diameter, which were evident in some of my initial visualizations of the data. KNN would be effective in capturing these localized relationships.

The results for the KNN regression are listed in the two tables below:

Variable	*Training*	*Testing*
Mean Squared Error	3.865191052315212	5.71881055004152
R^2 score	0.6151314091656529	0.5136750502669569

Table 3-3-A: The Training and Testing MSE and R^2 following the KNN Regression Model

Variable	*Importance*
Sex	0.061380
Length	0.029411
Diameter	0.025665
Height	0.011207
Whole Weight	0.847734
Shucked Weight	0.482224
Viscera Weight	0.018852
Shell Weight	0.240648

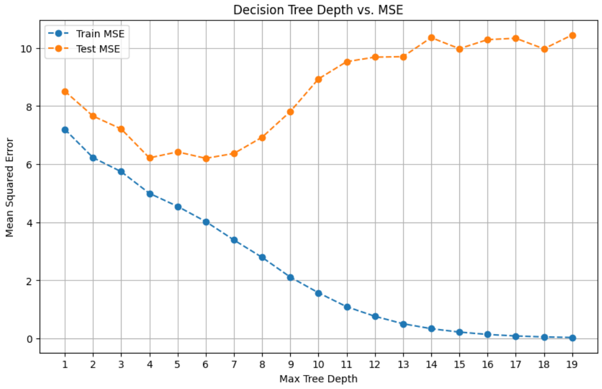
Table 3-3-B: Permutation Importance of Variables following the KNN Regression Model

Overall, both the training data and the testing data in my KNN model outperformed my baseline model, with the training data performing slightly better than the testing data. I believe this improvement is because the KNN model was able to account for variations in the data by utilizing the information within the independent variables to make predictions. The model leveraged the importance of features such as Whole Weight, Shucked Weight, and Shell Weight, which had the most significant impact on the predictions, to capture complex relationships in the dataset. This allowed the model to make more nuanced predictions compared to the baseline, which only predicted the mean abalone age and could not account for feature-specific variations.

The most significant input for the KNN Model was the whole weight of the abalones, followed by shucked weight. Height and viscera weight and diameter were the least important for predicting the number of rings in this model.

3-4. Decision Tree Regression Model

Next, I worked on the Decision Tree Regression Model. Decision trees are well-suited for capturing non-linear relationships within the abalone ring data. Their applicability in handling multicollinearity, which is a prevalent issue in the dataset, makes them an effective choice for this scenario. One of their key advantages is their clear and interpretable decision-making process, which makes it easy to understand how predictions are made based on feature values. This interpretability also provides valuable insights into the most important factors influencing abalone rings, helping to identify the best predictors for ring estimation.



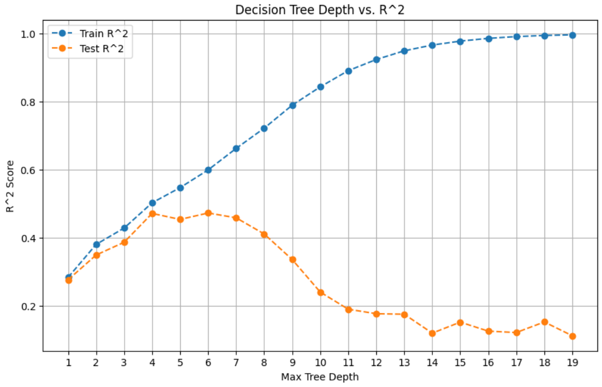


Fig. 3-4-A: Tree Depth for MSE
Fig. 3-4-B: Tree Depth for R^2

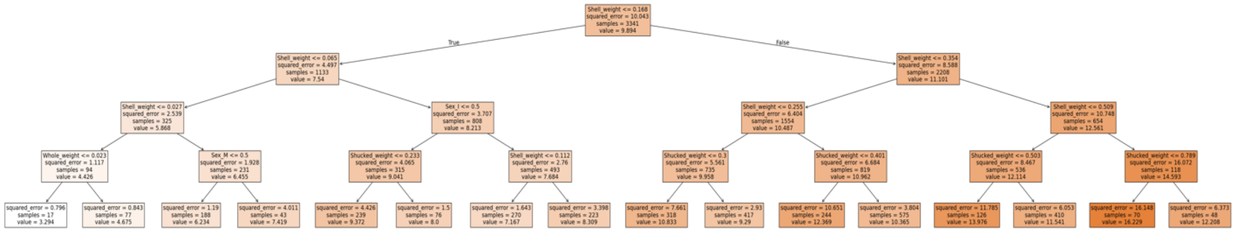


Fig. 3-4-C: Decision Tree

From the two graphs fitting tree depth respectively against main squared error and R^2, I found that when the max depth = 4, the train and test data perform best. The results for the decision tree model with a best max depth of 4 are listed in the two tables below:

Variable	*Training*	*Testing*
Mean Squared Error	4.996927547443125	6.217709768977742
R^2 score	0.5024410339215157	0.471248896183351

Table 3-4-A: The Training and Testing MSE and R^2 following the Decision Tree Regression Model

Variable	*Importance*
Sex_F	0.000000
Sex_M	0.004366
Sex_I	0.025921
Length	0.000000
Diameter	0.000000
Height	0.000000
Whole Weight	-0.000004
Shucked Weight	0.149245
Viscera Weight	0.000000
Shell Weight	1.124097

Table 3-3-B: Permutation Importance of Variables following the Decision Tree Regression Model

While my decision tree's training data outperformed my baseline regression, it performed slightly worse than my KNN model. This may be because the max depth of 4 was too low that, while reducing overfitting, it might not have been sufficient to capture the complexity of the underlying data patterns as effectively as the KNN model.

With this depth, the decision tree focused mostly on features like shell weight and shucked weight, considering them to be the two most important factors for predicting abalone rings. Other features, like whole weight, female, length, diameter, height, and viscera weight, were simply not used, as their permutation importance values were 0.

The use of the Decision Tree model is complex and controversial. On the one hand, it showed that weight-related features, such as shell weight and shucked weight, are key to predicting abalone rings, on the other hand, it did not use other features or capture more complex patterns. This may explain why its performance was worse than KNN.

3-5. Random Forest Regression Model

Extending the Decision Tree Model into a Random Forest Regression Model is a necessary step following the last one. Since my single decision tree did not perform very well, I wanted to explore a more comprehensive method, like random forest, for further improvement since they combine multiple trees to improve predictive performance.

Variable	*Training*	*Testing*
Mean Squared Error	1.6229800192605077	5.3368471723553865
R^2 score	0.8383950432176067	0.5461570356077239

Table 3-5-A: The Training and Testing MSE and R^2 following the Random Forest Regression Model

Variable	*Importance*
Sex	0.020121
Length	0.013126
Diameter	0.007149
Height	0.048467
Whole Weight	0.082206
Shucked Weight	0.624759
Viscera Weight	0.017836
Shell Weight	1.010396

Table 3-5-B: Permutation Importance of Variables following the Random Forest Regression Model

Overall, the Random Forest Regression Model outperformed all the previous three regression models, proving our expectations for it to further improve the reliability of the models is successful. Although the MSE for testing data is significantly higher than the training data, and the R^2 for testing data is significantly lower than the training data, they still outperformed the corresponding data in the other three models, proving that the Random Forest Model was the most successful one among all models that have been tested.

The permutation importance results of the Random Forest argued for the significance of shell weight and shucked weight. It also proves that diameter, length, and viscera weight were the least important inputs, which aligns with the findings in KNN and Decision Tree models.

3-6. Ridge Regression Model (A new model that was not introduced in class)

Finally, I wish to use the Ridge Regression Model and Test if it works for my dataset. Ridge regression makes predictions by minimizing the error between predicted and actual values while applying a penalty to large coefficients. If abalone ages were influenced by a combination of physical measurements, such as length, weight, or diameter, ridge regression would be effective in handling these relationships. The regularization term in ridge regression prevents overfitting by shrinking the impact of less important features. The results for the ridge regression are listed in the two tables below:

Variable	*Training*	*Testing*
Mean Squared Error	4.604362049152808	5.8050775482877235
R^2 score	0.5415299503792619	0.5063389454565954

Table 3-6-A: The Training and Testing MSE and R^2 following the Ridge Regression Model

Variable	*Importance*
Sex_M	0.097017
Sex_I	-0.791229
Length	2.342805
Diameter	7.049101
Height	6.753599
Whole Weight	7.046510
Shucked Weight	-17.085839
Viscera Weight	-6.795782
Shell Weight	10.046177

Table 3-6-B: Permutation Importance of Variables following the Ridge Regression Model

The regression results outperformed the Decision Tree Regression Model but were slightly inferior to KNN and Random Forest Models. A notable feature regarding the Ridge Model is that the difference in MSE and R^2 between the train and test sets is very small. This shows that the model balances bias and variance well and means the model performs reliably on unseen data. Ridge regression uses regularization to limit the impact of less important features. This helped the model avoid overfitting and achieve consistent results on both the training and testing datasets.

When it comes to the permutation importance, a key feature in the Ridge Model is that it exemplifies the coefficients, making the difference between them greater, which aligns with the different ways of handling relationships between Ridge and previous models. In the Ridge Model, the most important variables are Shell Weight, Diameter and Whole Weight, and Viscera Weight and Shucked Weight became less important, which are different from the results from previous models that emphasize shucked weight but ignoring diameters.

4. Discussion and Next Steps

4-1. Summary of Findings

In my analysis of abalone ring predictions, all the models I built performed better than the baseline predictor, showing their effectiveness and importance. The models, ranked by performance, are as follows: Random Forest Regression, K-Nearest Neighbors Regression, Ridge Regression, and Decision Tree Regression.

- Key Findings:

1) Superiority of the Random Forest Model: The Random Forest model was the most effective model among all the regression techniques I chose, which demonstrates that it had the best predictive capabilities and its best suitability for capturing complex relationships within the abalone ring dataset.

2) **Significant Features**: The most significant features within the Random Forest Model are Shell Weight and Shucked Weight. Shucked Weight was the most pivotal in all models except for the Ridge Regression model, indicating its high significance in the majority of the cases. Overall, weight-related factors are more important than length-related ones, as Whole Weight and Shell Weight were generally the most vital ones, while Viscera Weight was not very significant across cases.

In conclusion, the Random Forest Regression Model, with its inclusion of multiple regression trees, was the most suitable model for predicting the abalone rings. Categorizing the inputs into three types: sex, length-related, and weight-related, the model predicts that weight-related inputs generally are the most important ones, with Shucked Weight, Whole Weight, and Shell Weight holding respective significance across models despite certain minimal differences in the cases where they are most applicable. The findings provide a profound understanding of predicting abalone age based on the weights of respective features, offering valuable insights for future analyses and predictive modeling in future oceanographic studies.

4-2 Next Steps/Improvements

To improve the predictive performance of the models and enhance insights into the factors influencing abalone age, I would consider the following next steps and potential improvements:

1. Incorporate Environmental Data:

Adding environment-related data, such as water quality, food availability, and habitat conditions, could provide a better understanding of the external influences on abalone age and growth.

2. Increase Dataset Diversity:

Expand the dataset by including abalones from different regions or varying environmental conditions. A more diverse dataset could help improve the generalizability of the models.

3. Integration with Real-World Applications:

Collaborate with marine biologists and fisheries to validate the models with data on other marine animals and ensure the practical applicability of predictions for conservation or industry purposes. This can ensure a prediction result that is more broadly applicable to a wider range of marine species, making the significance of one model be applicable to the studies of different creatures.

By incorporating these improvements, the predictive models could be refined to provide more accurate estimates of abalone age and offer more actionable insights for marine biology and fisheries management. These steps would also help address potential gaps in the current analysis and open avenues for further research.