Abalone Age Prediction Using Lasso and Random Forest

1. Introduction

The age of abalone holds significant value for both abalone farming and the market. Typically, the age is determined by counting the number of rings, meaning the age of an abalone is equal to the number of rings plus 1.5. However, identifying the number of rings requires a biologist to cut the calcareous shell, stain it, and then observe and count the rings under a microscope to determine the abalone's age. This process is labor-intensive and costly. To streamline this process and improve efficiency, this report will explore the relationship between various physical measurements and the number of rings in abalone to predict the age of abalone. This study utilizes the abalone dataset from the UCI Machine Learning Repository (Nash, 1995), which contains 4177 observations.

Table 1 Variable description

Variable Name	Description	Units
Sex	M, F, and I (infant)	
Length	Longest shell measurement	mm
Diameter	perpendicular to length	mm
Height	with meat in shell	mm
Whole weight	whole abalone	grams
Shucked weight	weight of meat	grams
Viscera weight	gut weight (after bleeding)	grams
Shell weight	after being dried	grams
Rings	+1.5 gives the age in years	

As shown in Table 1, the abalone dataset includes nine variables: Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight, and Rings. In this study, Rings is considered the target variable, with the remaining variables serving as predictors. This study employs Lasso regression and Random Forest regression models to predict the number of abalone rings, thereby indicating the age of the abalone.

2． Data Cleaning and Exploratory Data Analysis

2.1 Data Cleaning

Firstly, I converted the type of predictor variable Sex to factors. After examination, no missing values were found in the dataset; however, outliers were present. Outliers can affect the training of the model and lead to instability. To address this, outliers in the data were detected and processed based on interquartile range (IQR). Values below the lower bound were replaced with the lower bound, and those above the upper bound were replaced with the upper bound. This Winsorizing approach not only reduces the negative impact of extreme values on the data analysis results, but also preserves most of the information in the retained data.

Table 2 provides a basic summary of all the variables, which is important for understanding the distribution of the data before modeling. For a more intuitive understanding, I visualized the relationship between the categorical variable Sex and the target variable Rings through two bar charts (Figure 1 and Figure 2). From Figure 1, It can be seen that the number of female and infant abalones is comparable, while males are higher than the other sex. Figure 2 reveals that the distribution of abalone rings is centrally clustered around 9.

Table 2 abalone data summary

Sex	Length	Diameter	Height	Whole_weight
F:1307	Min. :0.2025	Min. :0.1550	Min. :0.0400	Min. :0.0020
I:1342	1st Qu.:0.4500	1st Qu.:0.3500	1st Qu.:0.1150	1st Qu.:0.4415
M:1528	Median :0.5450	Median :0.4250	Median :0.1400	Median :0.7995
	Mean :0.5244	Mean :0.4083	Mean :0.1393	Mean :0.8274
	3rd Qu.:0.6150	3rd Qu.:0.4800	3rd Qu.:0.1650	3rd Qu.:1.1530
	Max. :0.8150	Max. :0.6500	Max. :0.2400	Max. :2.2203

Shucked_weight	Viscera_weight	Shell_weight	Rings
Min. :0.0010	Min. :0.0005	Min. :0.0015	Min. : 3.500
1st Qu.:0.1860	1st Qu.:0.0935	1st Qu.:0.1300	1st Qu.: 8.000
Median :0.3360	Median :0.1710	Median :0.2340	Median : 9.000
Mean :0.3578	Mean :0.1803	Mean :0.2380	Mean : 9.766
3rd Qu.:0.5020	3rd Qu.:0.2530	3rd Qu.:0.3290	3rd Qu.:11.000
Max. :0.9760	Max. :0.4923	Max. :0.6275	Max. :15.500

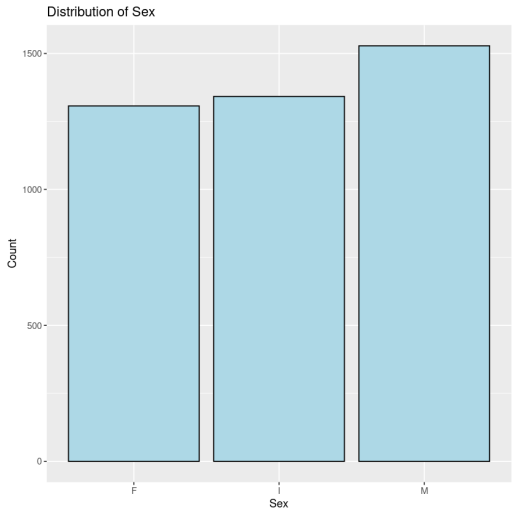

Figure 1. Distribution of Sex

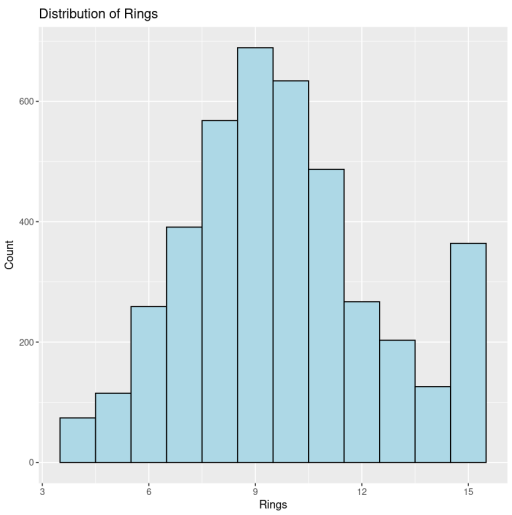

Figure 2. Distribution of Rings

The relationship between the variables was explored by plotting scatter plots (Figure 3), and the degree of correlation was further visualized through a correlation matrix and heatmap (Figure 4). Figures 3 and 4 indicate a significant linear relationship between 'Length' and 'Diameter', and a strong correlation among the four weight-related variables of abalone. This is in line with biological common sense, as abalones increase in size, both length and width also increase. Since the four weight variables contribute to the total body weight, their correlation is justified. For the target variable 'Rings', while the relationship with other variables is not obvious, it can still be judged to have a largely positive correlation with the other variables, suggesting that as the abalone grows in age, these measurements will largely grow as well. Overall, the predictor variables in this dataset show strong intercorrelations, and they are largely positively correlated with the response variable.

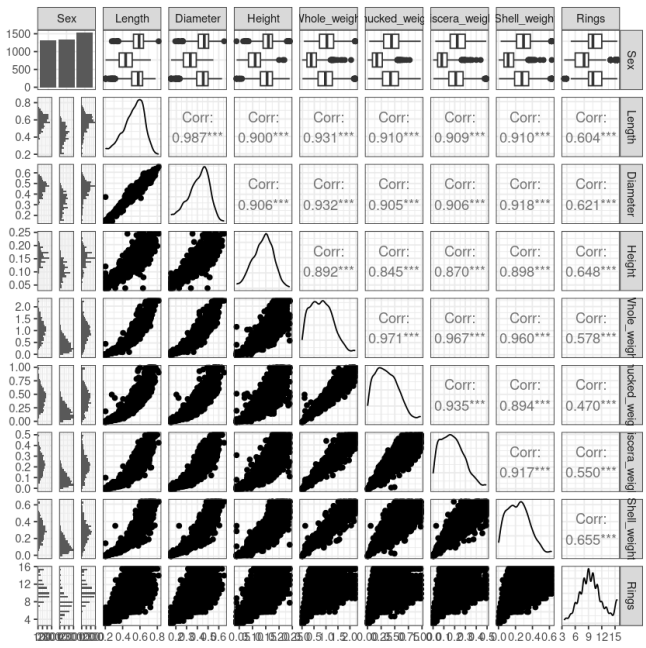

Figure3 Scatterplot of abalone variable relationship

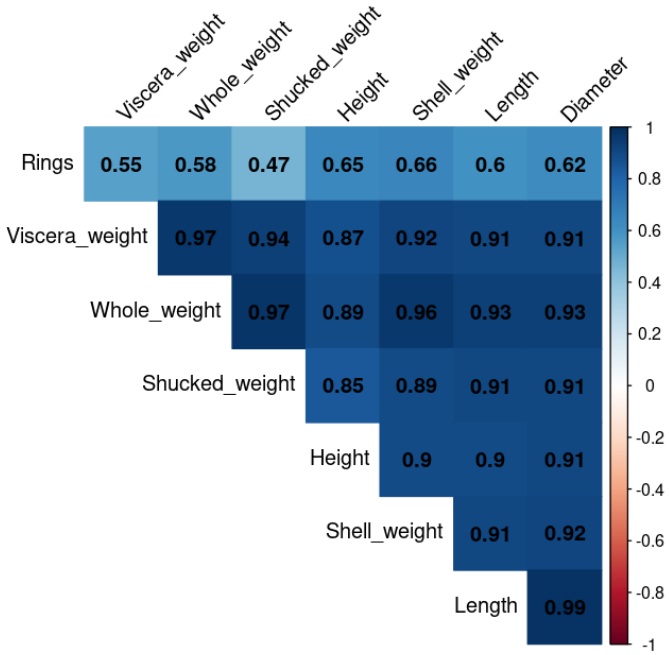

Figure 4 Heat map representing Data Correlation

Prior to modeling, the categorical variable Sex was one-hot encoded as shown in Table 3, creating a dummy variable for each gender category. This allows each gender level to be interpreted in the model based on its contribution to the target variable Rings. The dataset was then divided into two parts, where 80% was used as a training set and 20% as a test set.

Table 3. Abalone data after cleaning

	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings	SexF	SexI	SexM
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
1	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15	0	0	1
2	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7	0	0	1
3	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9	1	0	0
4	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10	0	0	1
5	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7	0	1	0
6	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.120	8	0	1	0

3. Modelling

The objective of the modeling phase was to construct models for predicting the age of abalone (expressed as the number of rings) using Lasso Regression and Random Forest. This section provides a brief explanation of each method and explores the results derived from each model.

3.1 Lasso Regression

3.1.1 Overview and Assumption

Lasso regression is a type of linear regression that incorporates a penalty term. In practical scenarios, there may be a large number of independent variables, some of which may possess weak or redundant predictive power for the target variable. Lasso regression addresses this issue by introducing L1 regularization (the Lasso penalty term), which facilitates feature selection and model sparsity. This regularization technique can adjust the weights of less significant elements in the coefficient vectors to zero, effectively reducing the complexity of the model. Within the context of the abalone dataset, applying Lasso regression implies the assumption of a linear relationship between the physical measurements of abalones and their age.

3.1.2 Results

For the abalone dataset, Lasso regression was used to determine which physical measurements best predicted the age of abalones. Initially, cross-validation was employed to construct the Lasso regression model by plotting the Mean Square Error

against the penalty factor lambda (Figure 5). The optimal lambda value corresponds to the minimum Mean Square Error, and this value was used to fit the Lasso model. Finally, the prediction is made and the MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), R-squared are 1.3488, 1.7238 and 0.6213 respectively.

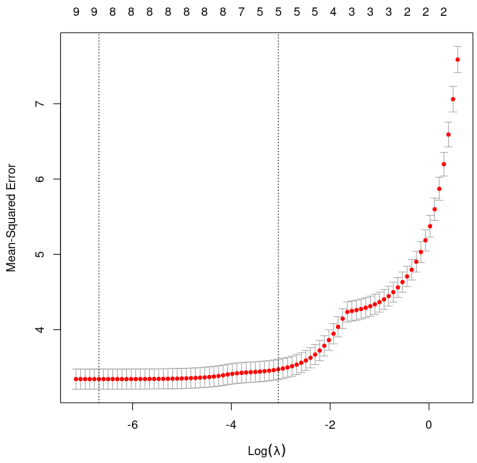

Figure 5. MSE changes of LASSO regression

	s1
(Intercept)	3.17523938
Length	0.05676502
Diameter	10.53475827
Height	20.40362628
Whole_weight	5.59315681
Shucked_weight	-15.43436151
Viscera_weight	-6.66000639
Shell_weight	7.19840044
SexF	.
SexI	-0.75814907
SexM	0.05623452

Figure 6. Coefficients of predictor variables

Figure 6 displays the correlation coefficients of each predictor variable in the Lasso regression model under the optimal lambda. The results show that the variables Diameter, Height, and Shucked weight have significant coefficients, indicating a strong linear relationship with Rings. The correlation coefficients for SexF, SexM, and Length are nearly zero, suggesting that their effects on the number of rings are not significant in this model.

3.2 Random Forest

3.2.1 Overview and Assumption

Random Forest is an integrated learning method that produces the final output by constructing multiple decision trees during training and aggregating the outputs of individual trees (Figure 7). It does not assume linearity in the data and can model complex interactions between features. We thus assume that the abalone dataset contains complex patterns that may not be captured by a linear model, and that these patterns can be better represented by a Random Forest.

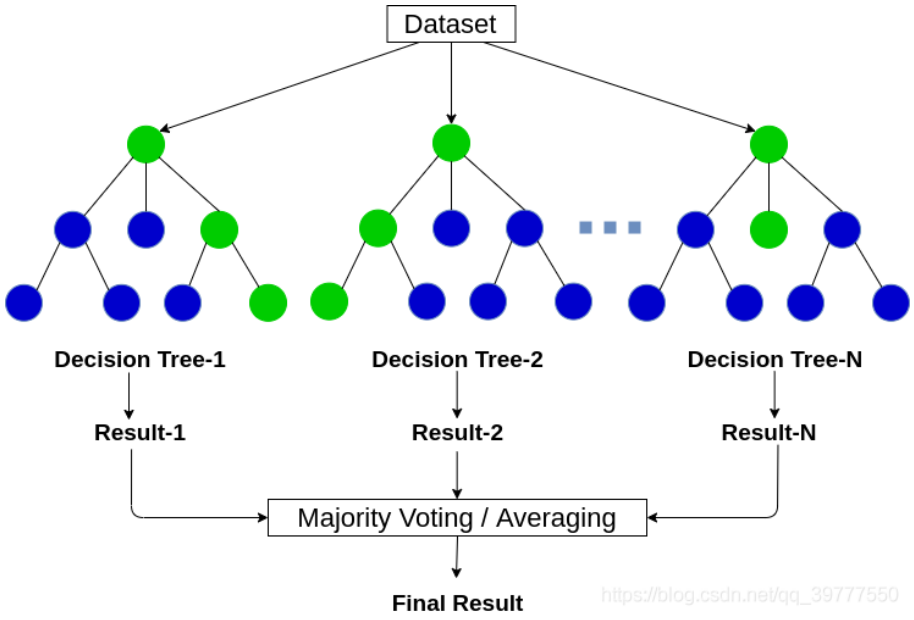

Figure 7. Random Forest Schematic (The relationship between decision trees and random forests, 2020)

3.2.2 Results

In constructing the Random Forest model, one must compare the Mean of Squared Residuals by manually adjusting the hyperparameters to obtain different 'ntree' and 'mtry' values. The optimal values I found were 800 for 'ntree' and 3 for 'mtry'. The model was then used for prediction, and a scatter plot (Figure 8) was generated to compare the predicted values with the actual values. The plot shows that the Random Forest model has reasonable prediction accuracy, as many points are close to the prediction line. However, numerous points deviate from the line, indicating potential for improving the model's predictive power. The MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R-squared were determined to be 1.3069, 1.6822, and 0.6415, respectively.

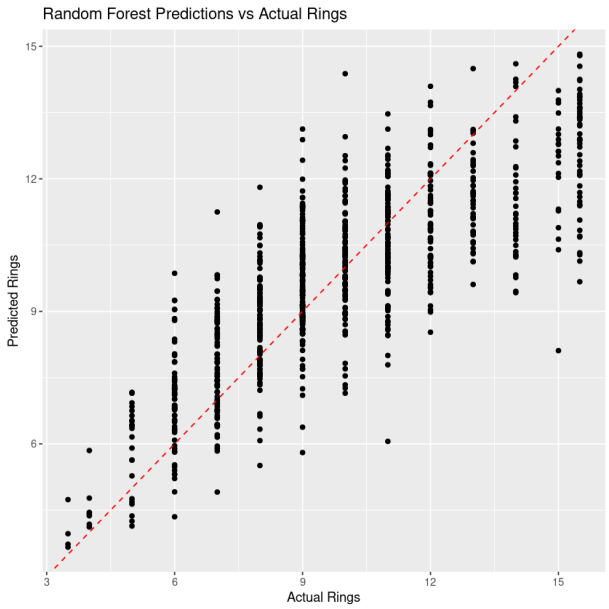

Figure 8. Random Forest Prediction vs Actual Rings

Figure 9 illustrates the importance of each predictor variable in the Random Forest model. The model identified Shell weight as the most influential feature, followed by Height, highlighting their significance in predicting Rings. Conversely, the three variables representing sex are less important in predicting Rings.

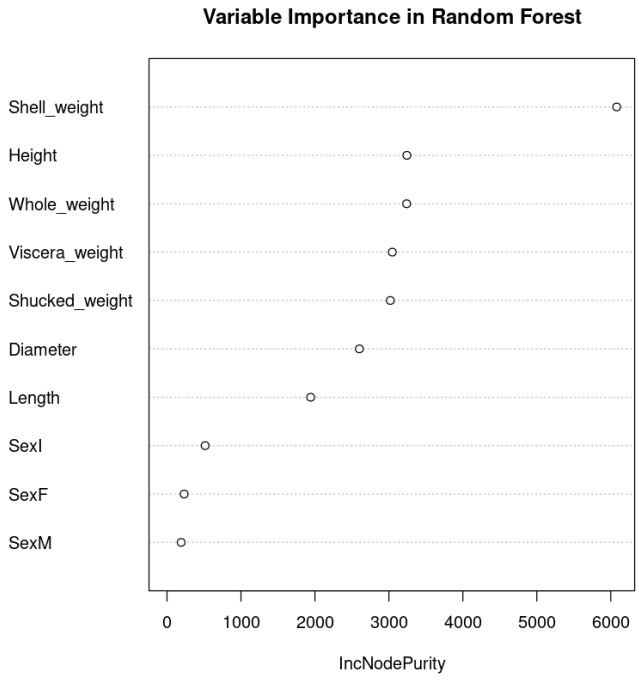

Figure 9. Variable Importance in Random Forest

4. Model Comparison

In this section, three performance metrics will be used to compare and assess the predictive accuracy and quality of fit of the two models.

4.1 Mean Absolute Error (MAE)

MAE is defined as the absolute average difference between each predicted values and observed values (Smola, 2008). It provides a direct measure of predictive accuracy on the same scale as the data. The lower the MAE, the higher the predictive accuracy of the model. A MAE of zero means that the model perfectly predicts all observations, though this is rarely the case in practice. Here is the formula for MAE:

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=i}^{n} \left|y_j - \hat{y}_j\right|$$

By comparing the MAE of the two models (Figure 10), it is evident that the MAE of Random Forest is lower than that of Lasso regression, indicating that the Random Forest model has relatively higher predictive accuracy.

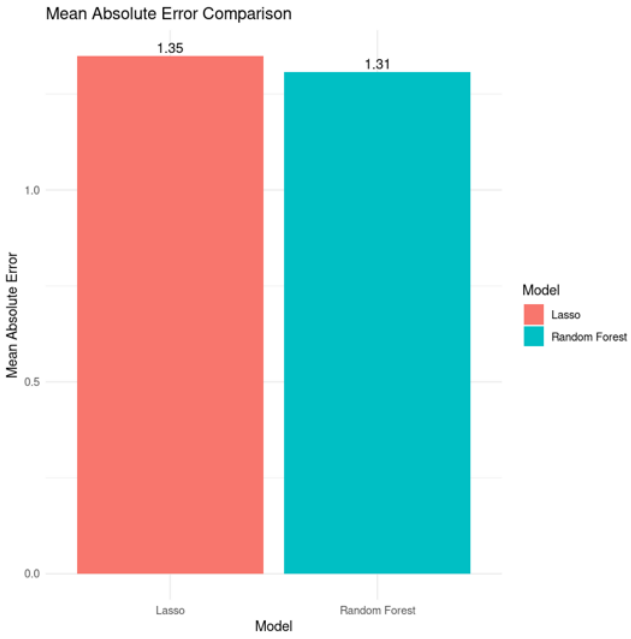

Figure 10. MAE Comparison

4.2 Root Mean Squared Error (RMSE)

RMSE calculates the square root of the squared average difference between each predicted values and observed values (Smola, 2008). Like MAE, a value of zero indicates a perfect fit. However, RMSE gives more weight to larger errors. This means that RMSE is sensitive to outliers and is more useful when large errors are particularly undesirable. Here is the formula for RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=i}^{n}(y_j - \hat{y}_j)^2}$$

Figure 11 shows that the RMSE for Random Forest is also lower than that for Lasso regression, suggesting that the Random Forest model has higher predictive accuracy.

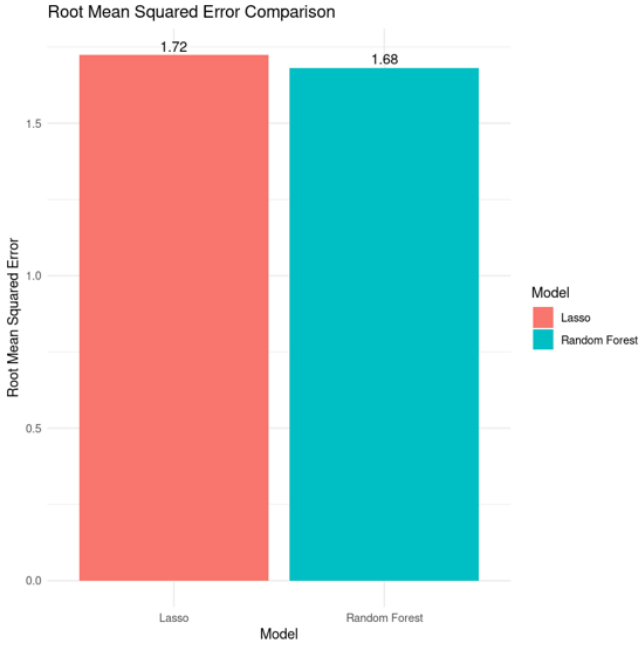

Figure 11. RMSE Comparison

4.3 R-squared (R^2):

R^2 reflects how well the model fits the data, with values generally ranging from 0 to 1. The closer the value of R^2 is to 1, the better the variables in the model explain the dependent variable, and the better the model fits the data (Indicators for the evaluation of regression models , 2022). Here is the formula for R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Figure 12 compares the R^2 values of the two models. It can be seen that Random Forest has a higher R^2 than Lasso, indicating that the Random Forest model better fits the data.

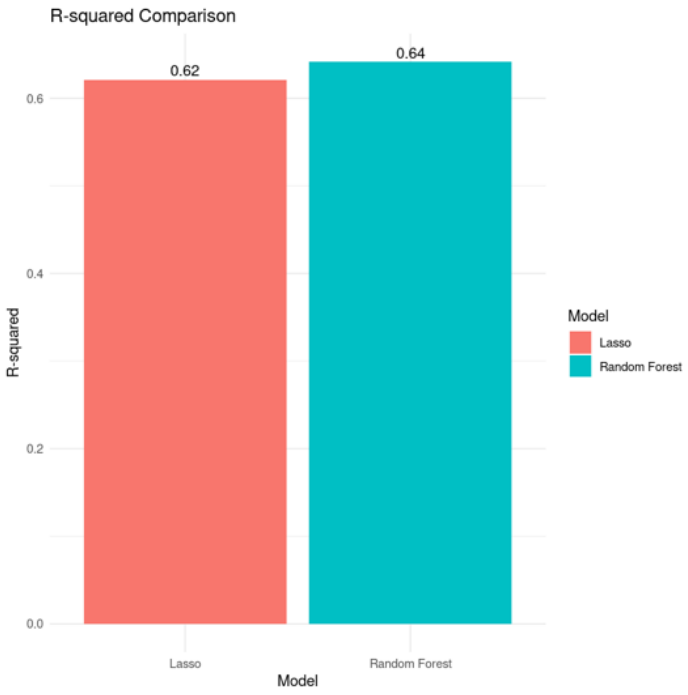

Figure 12. R^2 Comparison

In conclusion, Random Forest exhibits higher overall accuracy than the Lasso model, with lower mean absolute and root mean square errors. Additionally, the Random Forest model has a higher R-squared value, indicating more explanatory power for the Rings variable. While the simplicity and interpretability of the Lasso model are its strengths, Random Forest demonstrates superior performance across all three metrics and may be better suited for this forecasting task.

5. Results and Conclusion

Based on the abalone dataset from the UCI Machine Learning Repository, this study establishes Lasso regression and Random Forest models using samples from the training set, and then predicts the rings of abalone using the test set. The advantages and disadvantages of the two models are compared through three performance

metrics, leading to the conclusion that the Random Forest model exhibits the best prediction performance. To reduce the tediousness of observing and calculating the number of abalone rings using traditional methods, the study employs a hyperparameter-adjusted Random Forest model. This model requires only the physical measurements of abalone to predict the number of rings, and thus the age, with a relatively high degree of accuracy. The limitations of this study include the presence of numerous predictor variables and the absence of feature selection to reduce computational complexity. Additionally, there is room for further improvement in the fit of both models, which at this stage, failed to make very accurate predictions. Future work could enhance predictions with more advanced regression techniques, such as neural networks.

6. Bibliography

Indicators for the evaluation of regression models . (2022, 4 5). Retrieved from CSDN: https://blog.csdn.net/y15659037739l/article/details/123971286

Nash, W. S. (1995). *Abalone. UCI Machine Learning Repository.* Retrieved from https://doi.org/10.24432/C55C7W

Smola, A. a. (2008). *Introduction To Machine Learning.* The Press Syndicate Of The University Of Cambridge.

The relationship between decision trees and random forests. (2020, 7 13). Retrieved from CSDN: https://blog.csdn.net/qq_39777550/article/details/107312048