Statistical Modeling Of Data Using R

Final-term project

Applied Statistics and Visualization for Analytics –515

Tarun Komirishetty
Sanisha Kolanu
Manoj Kudikala

Group- 10

George Mason University College of Engineering and Computing
Data Analytics Engineering Program
Fall 2022

Instructor: Isuru Dassanayake, PhD.

Abstract:

This report deals with the Abalone dataset where we fit different data models to predict the age of the Abalone by using Rings attribute as the response variable and all other attributes like sex, length, Diameter, Height as the predictor Variables. We also predict the sex of the abalone as a categorical response variable using the other attributes as predictors. The first data model we do is best Subset selection and then K-fold Cross Validation, tree, Random Forest methodologies are applied on the given data set divided into test and train parts to obtain the response variable and measure its accuracy to find the best fit model for the considered Abalone dataset.

Problem Statement

Predicting the age of abalone from the physical measurements. The age of the abalone is determined by cutting the shell through the cone and staining it to count the number of rings and add 1.5 to it for determining the age of the Abalone which is a lot of time-consuming task. So, we need to determine the best model for finding the age of the Abalone.

Research Questions:

I)To determine the Age of the Abalone based on the Number of Rings in the dataset.

II)To determine the ratio of the sex whether its Male, Female or Infant In the given Dataset observations.

III)Finding the best model to predict the Age of the Abalone using MSE.

Dataset Description:

Dataset Description: The dataset has about 4177 observations with 9 variables. The attribute in the dataset is divided into one categorical and eight continuous variables. There are 3 different genders (M, F, and I). The number of rings which is the response variables can be estimated by the predictor variables. We can also estimate the gender based on the other predictor variables.

Attributes in the dataset:

VARIABLE	TYPE	DESCRIPTION
Sex	Categorical	Sex of the snail (M, F, I)
Length	Continuous	Length of the shell in mm
Diameter	Continuous	Diameter perpendicular to the length in mm
Height	Continuous	Height of the abalone in mm
Whole Weight	Continuous	Weight of the whole Abalone

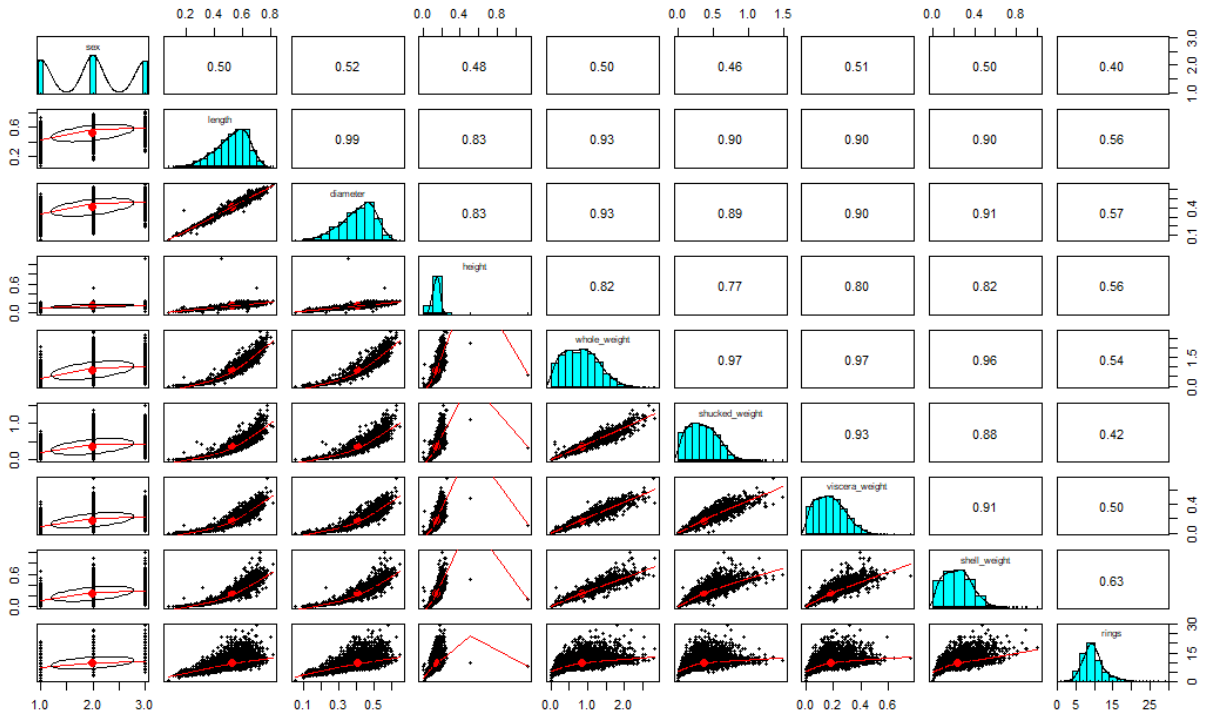
Shucked Weight	Continuous	Weight of the meat of the Abalone
Viscera Weight	Continuous	Gut weight of the abalone
Shell Weight	Continuous	Weight of the abalone after being dried
Rings	Continuous	Number of rings which gives the age of the abalone

MODELING

The problem we are trying to solve is to estimate the number of rings of an abalone and sex. The modeling algorithms we used are Best Subset selection, K Fold Cross Validation, Regression tree, Classification Tree, and Random Forest. We looked at Mean Squared Error value to compare which model is best.

1. Correlation plot

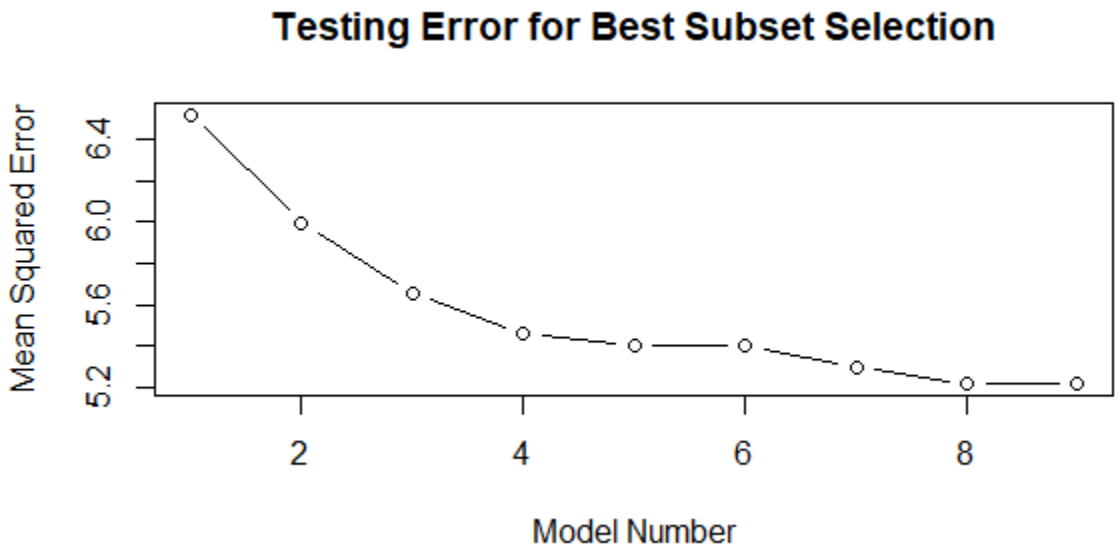
To see the descriptive statistics of the attributes, we use pairs.panels to see the scatter plot matrix. Correlation between the attributes can also see in the below figure. From the below plot, there is high positive correlation of 0.63 between shell weight and number of rings. We can see that there is a linear relationship between number of rings and whole weight, shucked weight, viscera weight, and shell weight.



I) Rings as the Response Variable

2. Best Subset Selection:

In the best subset selection, we try to reduce the number of variables used to predict the response variable to get a less complex model. We used the regsubsets function in leaps package to perform the best subset selection with number of rings as the response variable and the others as predictors. First, we divide the data into testing and training datasets with 60% of the observations in training data and 40% of the observations in the testing dataset. We try to determine the mean squared error value of the testing dataset. Since we have 8 predictor variables with a categorical variable of 3 levels, we come up with 9 models. We then calculate the testing mean squared error for all the 9 models and select the model with the least mean squared error. The graph of Mean squared error for the 9 models is given below.



We can see from the above graph that 8th model has the least mean squared error. The coefficients of the 8th model are given below.

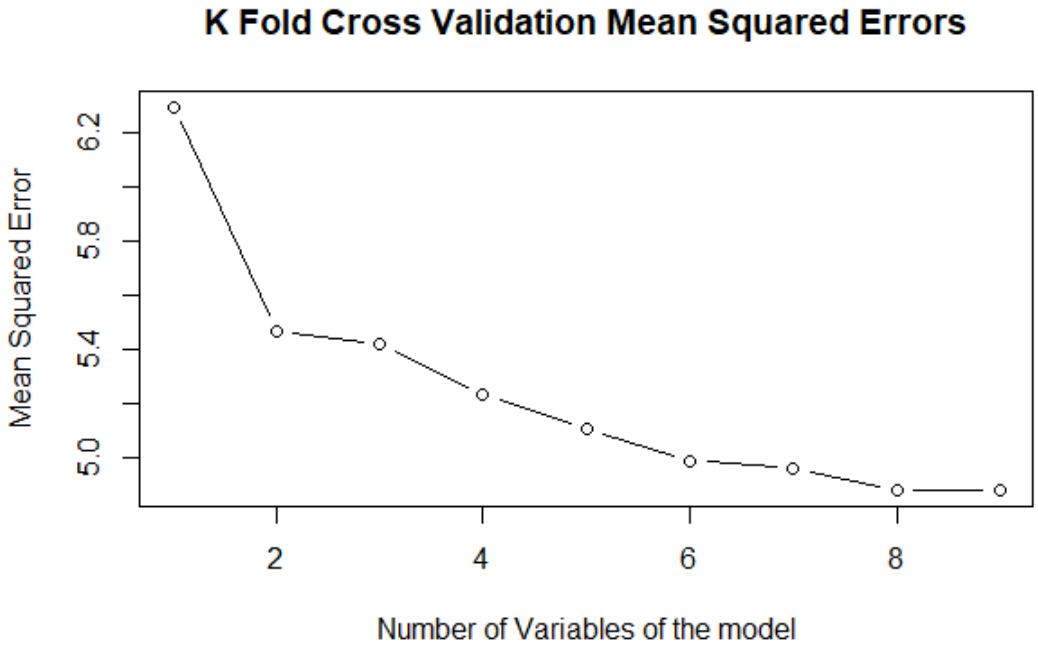
Best Subset Selection	
Variables	Estimates
Intercept	3.180
sex1	0.910
sex2	0.914
diameter	11.032
height	6.783
whole weight	10.088
shucked weight	-21.490
viscera weight	-10.473

Shell Weight	8.066

The minimum value of mean squared error is 5.2166.

3. K Fold Cross Validation:

We have performed a 10-fold cross validation where we have divided the whole dataset into 10 parts and each observation gets to be in the training set and the testing set. In this way we can eliminate the division of observations into training and testing data. We fit 9 models for each of the 10 sets of data divisions and come up with 90 different models. We calculate the mean squared error for all the models using the testing datasets. Then we find which model has the least mean squared error. The below graph shows the number of variables with which we get the least mean squared error.



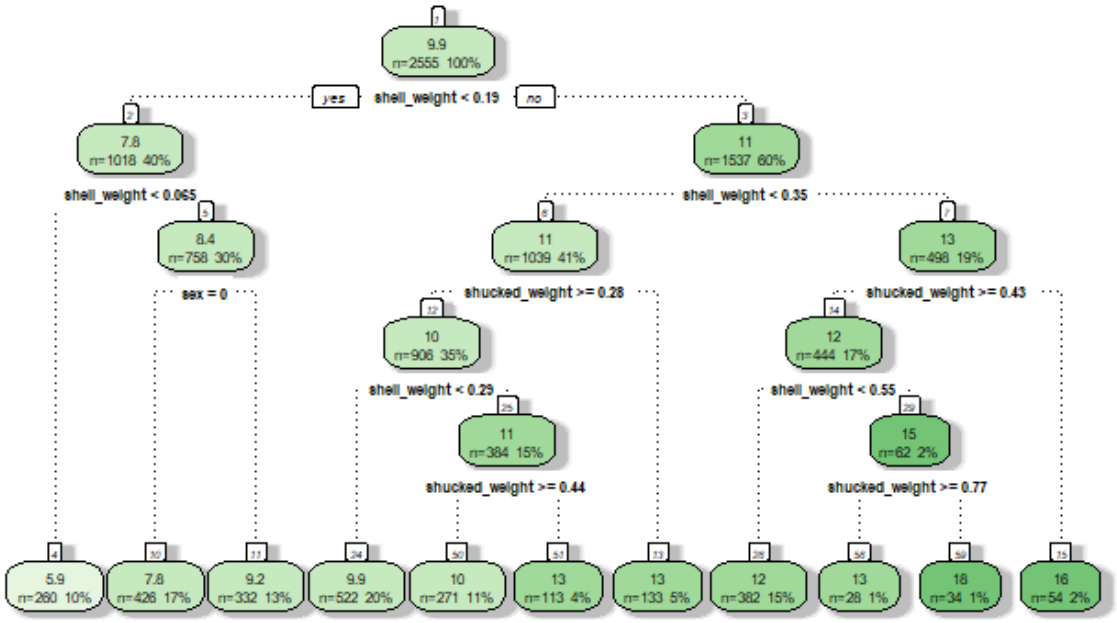
We can see that the model with 8 variables has the least mean squared error. The co-efficient of the 8th model are given in the table below. The MSE value is 4.1273.

K Fold Cross Validation	
Variables	Estimates
Intercept	3.043
sex1	0.883
sex2	0.826
diameter	10.569
height	10.749
whole weight	8.997

shucked weight	-19.802
viscera weight	-10.612
Shell Weight	8.750

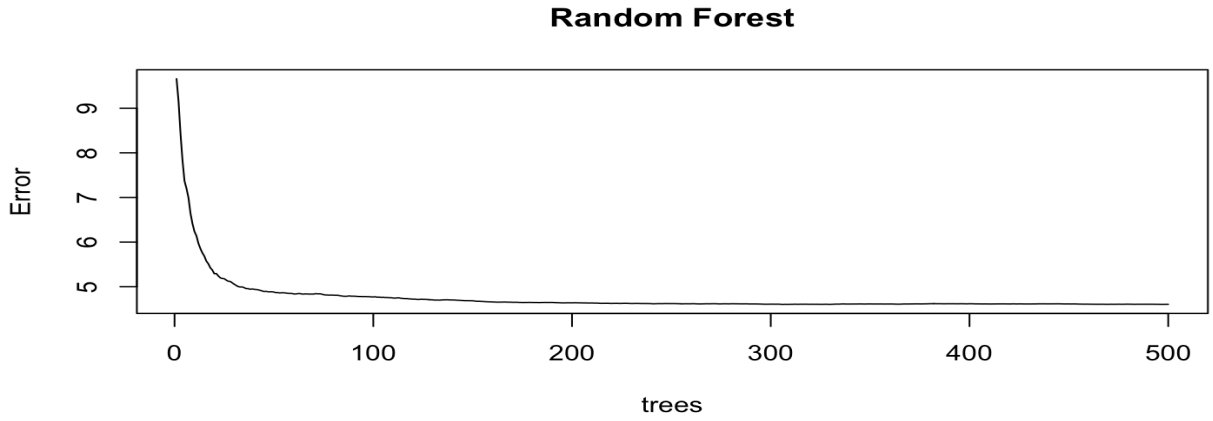
4. Tree Method:

We have applied tree model to determine the number of rings of the abalone. The tree makes the recursive partitioning of the data to minimize the residual deviance by splitting the data into two subsets. The p-values or the gini coefficient methods are used to select and split the variables.The variables used in the tree construction were “shell weight”, “sex” and “shucked weight”. There are 11 terminal nodes in the tree. The residual mean deviance of the tree model is 5.101.



5.Random Forest:

In this method, we are modeling 500 Trees with three variables each by minimizing the mean squared error rate of the test dataset to predict the values with utmost accuracy. The MSE value of the random forest model is 4.697



Visualizations:

Future Work

The Future analysis can be performed by consulting experts in the area of the domain addition of additional features like geographic location of abalone, color of abalone ,their species, living environment, etc. for more accurate analysis.

Limitations:

One Important founding regarding the limitation is that there is the high correlation between total weight and length of Abalone which is 0.97 indicating these two attributes are positively correlated. As a result, it can become difficult for the model to estimate the relationship between dependent variable and each independent variable independently.

We fit best subset selection,Random forest,tree,kfold models for our dataset by considering their performance but further work can be done on feature selection or best model selection to increase the performance.

There are some unnecessary features category observations in the project dataset like Infant sex of abalone which may harm the data validity.

The lack of domain feature analysis for model inputs is the biggest limitation to the project which leads to the addition of all the attributes of the dataset for predicting Number of Rings. However, if greater knowledge is achieved by consulting experts, we can do feature selection more correctly and increase the reliability and performance of model.

Conclusion:

After fitting different models like Bestsubset selection,Kfold CrossValidation,Tree,Randomforest on the Abalone database we can see that Random Forest method is the best model with utmost accuracy for prediction of the Age of Abalone .