PH125.9x Capstone Abalone Project

Michael W. Jones

8/19/2020

Introduction

Executive Summary

Abalone, any of several marine snails, constituting the genus Haliotis and family Haliotidae in the subclass Prosobranchia (class Gastropoda), in which the shell has a row of holes on its outer surface. Abalones are found in warm seas worldwide.

The dishlike shell is perforated near one edge by a single row of small holes that become progressively filled during the animal's growth; the last five to nine holes remain open to serve as outlets for the snail's waste products.

The shell's lustrous, iridescent interior is used in the manufacture of ornaments. The large muscular foot of the abalone is eaten as a delicacy in several countries.

Depending on the species, abalones usually range from 10 to 25 cm (4 to 10 inches) across and up to 7.5 cm in depth. About 50 species have been described. The largest abalone is the 30-cm red abalone (H. rufescens) of the western coast of the United States. H. rufescens and several other species are raised commercially in abalone farms, particularly in Australia, China, Japan, and along the western coast of the United States. Commercial fisheries for abalones exist in California, Mexico, Japan, and South Africa. (Britannica, 2020)

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope - a tedious and time-consuming task. Other measurements, which are easier to obtain, may be used to predict the age. The age is calculated as 1.5 plus the number of rings.

Additional information, such as weather patterns and location (hence food availability) may be required to accurately predict age (i.e. rings).

Overview

This project uses the information from the abalone dataset in the PivotalR library to identify the best model for predicting rings.

Dataset Format

The attribute name, attribute type, the measurement unit and a brief description is given. The number of rings is the value to predict: either as a continuous value or as a classification problem.

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

Key Steps

The abalone dataset is loaded from the PivotalR library. It is then divided into training and testing data sets, 90% and 10% respectively.

In order to determine the best model for predicting rings, and hence the highest accuracy, both regression and classification analyses are done. Models used for regression are: lm, glm, knn, svmLinear, rpart, gamLoess, treebag, gbm, and rf. The classification models are: knn, lda, qda, naive_bayes, svmLinear, svmRadial, gamLoess, multinom, and rf. Classification is done by grouping rings into age groups: juvenile, adult, and senior. The best models (least RMSE / highest accuracy) are chosen from each set of analysis. The results and conclusions from the analysis are reported.

Data Prep

The abalone dataset is obtained from the PivotalR library. Sex is converted to a number: I = 0; F = 1; and M = 2 for both regression and classification analysis. For classification, rings are grouped as juvenile (1-7); adult (8-11); and senior (12-29). The grouping was based on a summary() of rings using the 1st quartile for juvenile, between the 1st and 3rd quartile for adult, and senior the remainder.

```
summary(abalone$rings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   8.000   9.000   9.934  11.000  29.000
```

Analysis Methods

Data analysis begins by visualizing the data. The abalone data is visualized in several different ways, including: looking at the actual data, the data structure, summary, correlation, density plots, histograms, box plots, pairs plot, and correlation plot.

Subsequently, regression analysis is performed using nine different model methods (lm, glm, knn, svmLinear, rpart, gamLoess, treebag, gbm, and rf). After grouping the rings as juvenile, adult, and senior, classification analysis is done using nine model methods as well (knn, lda, qda, naive_bayes, svmLinear, svmRadial, gamLoess, multinom, and rf).

The results of the analyses are reported for each analysis type - regression and classification, identifying the best model for each.

Visualizing the data

Data Sample

```
head(abalone, 3)
```

```
##   sex length diameter height  whole shucked viscera shell rings
## 1   M  0.455    0.365  0.095 0.5140  0.2245  0.1010  0.15    15
## 2   M  0.350    0.265  0.090 0.2255  0.0995  0.0485  0.07     7
## 3   F  0.530    0.420  0.135 0.6770  0.2565  0.1415  0.21     9
```

Structure

```
str(abalone)
```

```
## 'data.frame':    4177 obs. of  9 variables:
##  $ sex     : chr  "M" "M" "F" "M" ...
##  $ length  : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
##  $ diameter: num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
##  $ height  : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
##  $ whole   : num  0.514 0.226 0.677 0.516 0.205 ...
##  $ shucked : num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
##  $ viscera : num  0.101 0.0485 0.1415 0.114 0.0395 ...
##  $ shell   : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
##  $ rings   : int  15 7 9 10 7 8 20 16 9 19 ...
```

Summary

```
summary(abalone_r)
```

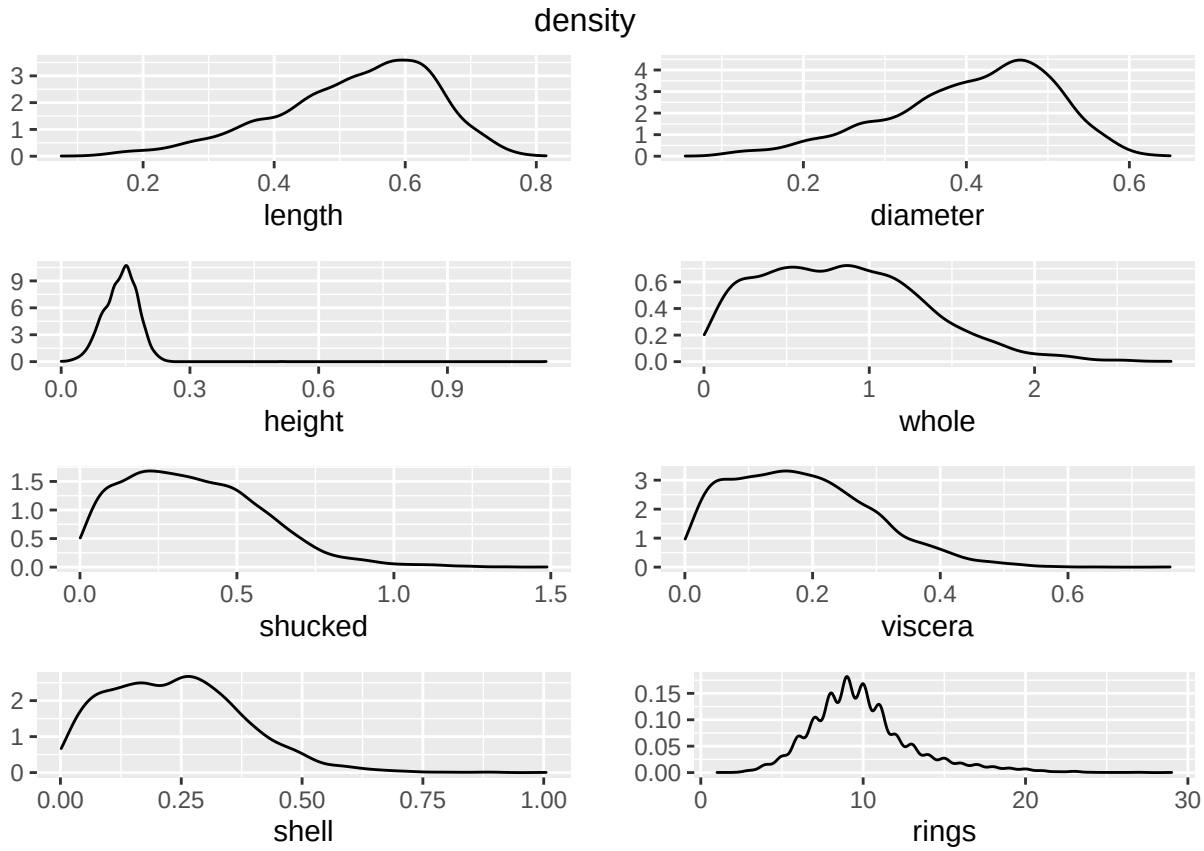
```
##     sex_num          length         diameter          height
##  Min.   :0.000   Min.   :0.075   Min.   :0.0550   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150
##  Median :1.000   Median :0.545   Median :0.4250   Median :0.1400
##  Mean   :1.045   Mean   :0.524   Mean   :0.4079   Mean   :0.1395
##  3rd Qu.:2.000   3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650
##  Max.   :2.000   Max.   :0.815   Max.   :0.6500   Max.   :1.1300
##      whole           shucked          viscera           shell
##  Min.   :0.0020   Min.   :0.0010   Min.   :0.0005   Min.   :0.0015
##  1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300
##  Median :0.7995   Median :0.3360   Median :0.1710   Median :0.2340
##  Mean   :0.8287   Mean   :0.3594   Mean   :0.1806   Mean   :0.2388
##  3rd Qu.:1.1530   3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290
##  Max.   :2.8255   Max.   :1.4880   Max.   :0.7600   Max.   :1.0050
##      rings
##  Min.   : 1.000
##  1st Qu.: 8.000
##  Median : 9.000
##  Mean   : 9.934
##  3rd Qu.:11.000
##  Max.   :29.000
```

Correlation data

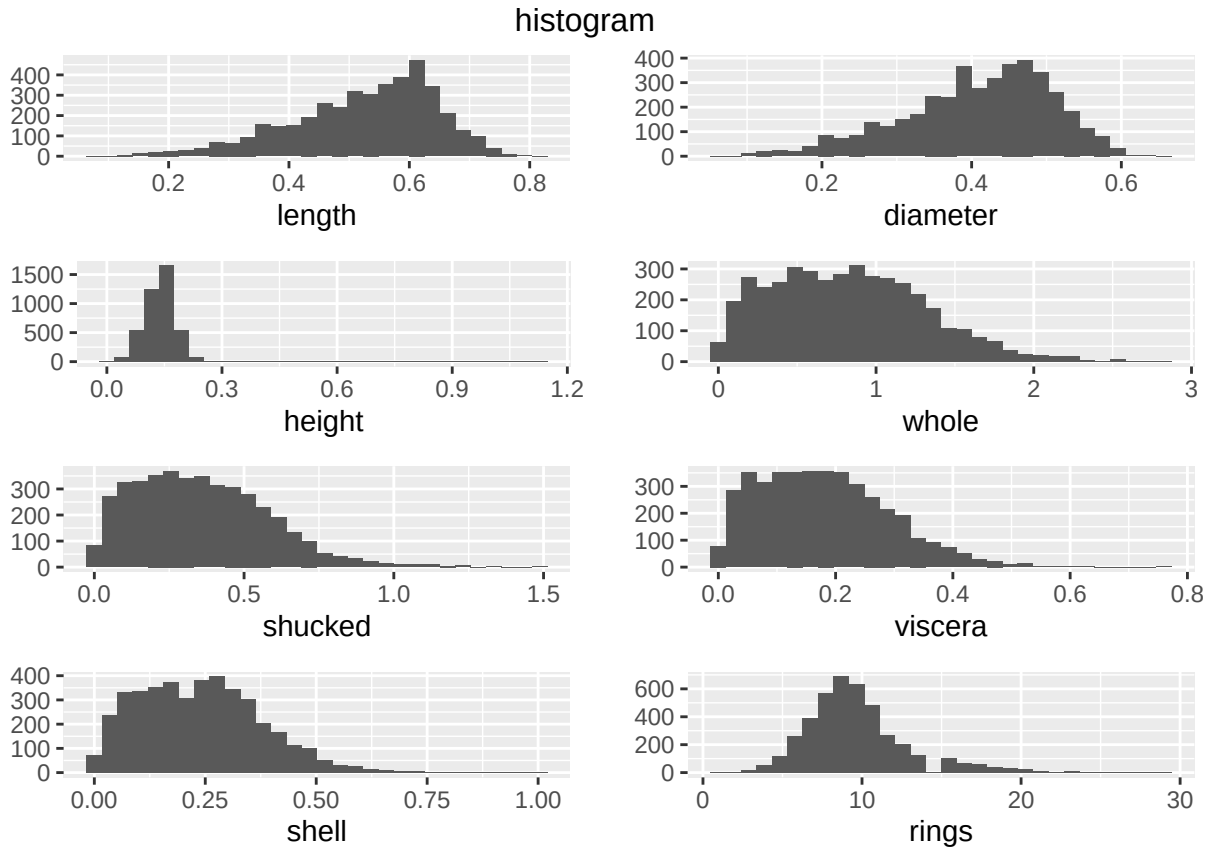
```
cor(abalone_r[9], abalone_r[-9])
```

```
##         sex_num    length  diameter    height     whole   shucked   viscera
## rings 0.3518216 0.5567196 0.5746599 0.5574673 0.5403897 0.4208837 0.5038192
##          shell
## rings 0.627574
```

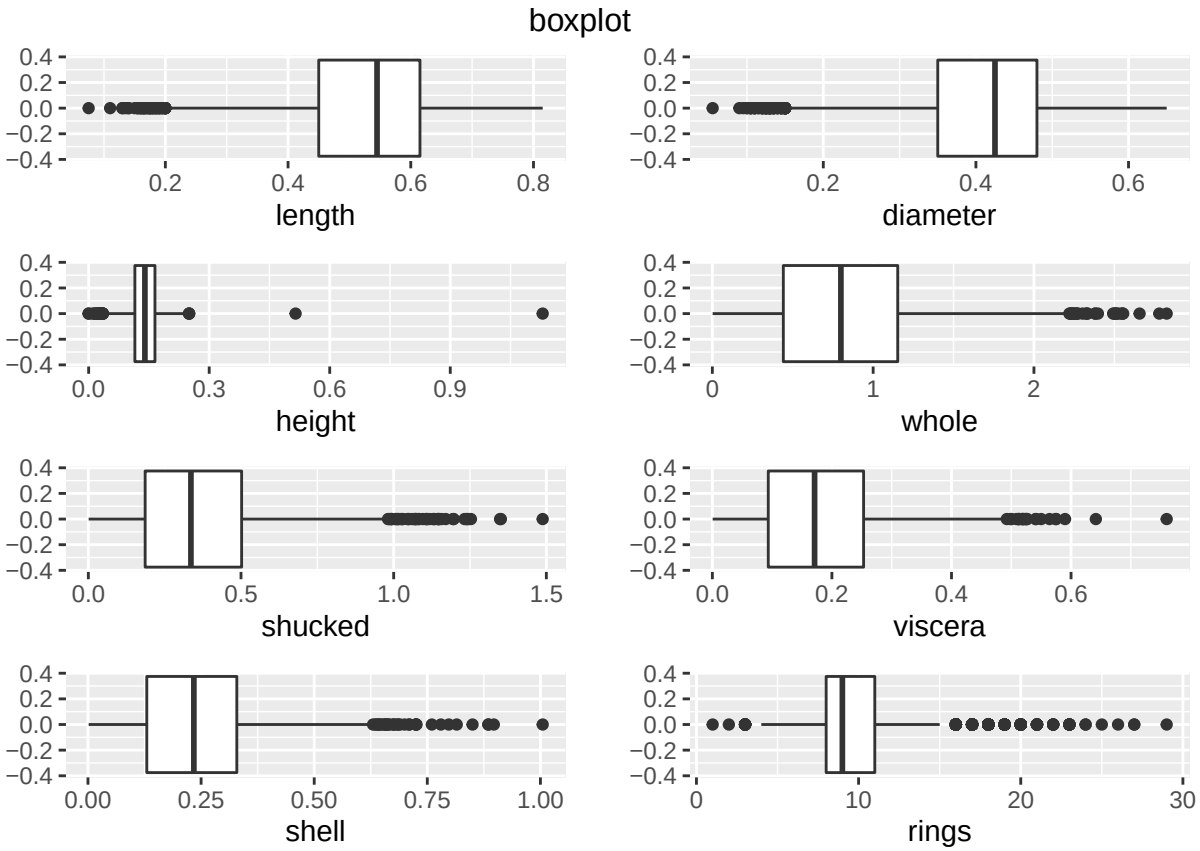
Density plots



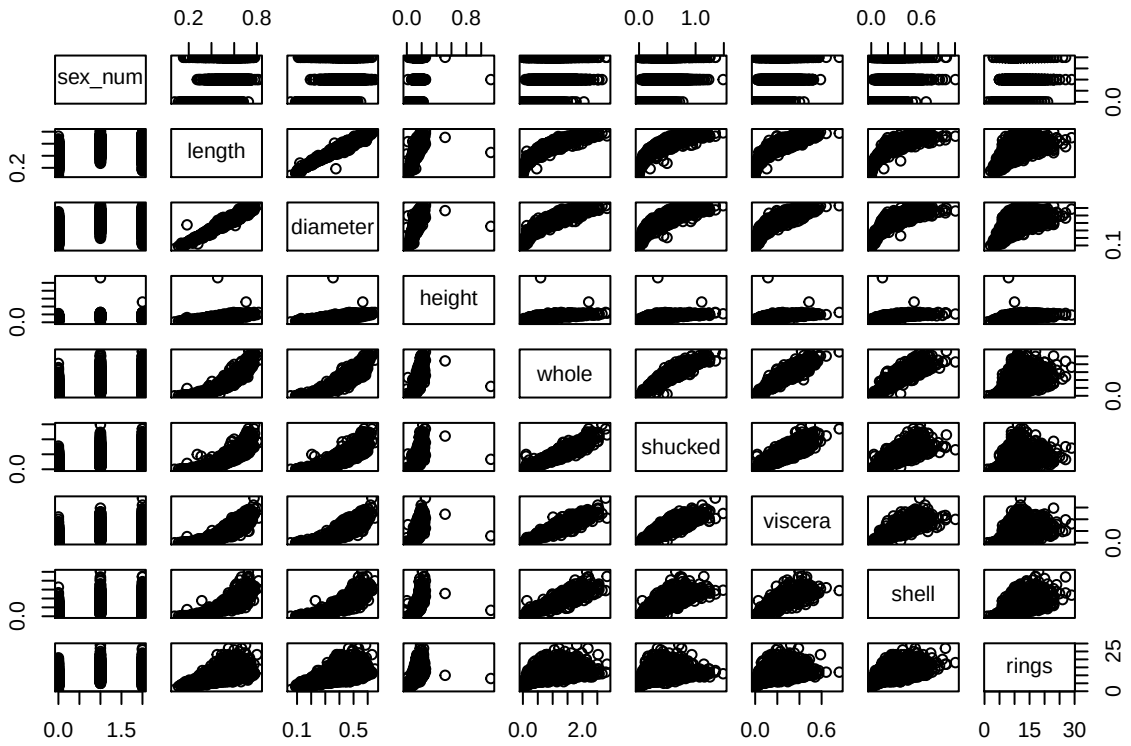
Histograms


histogram
length
diameter
height
whole
shucked
viscera
shell
rings

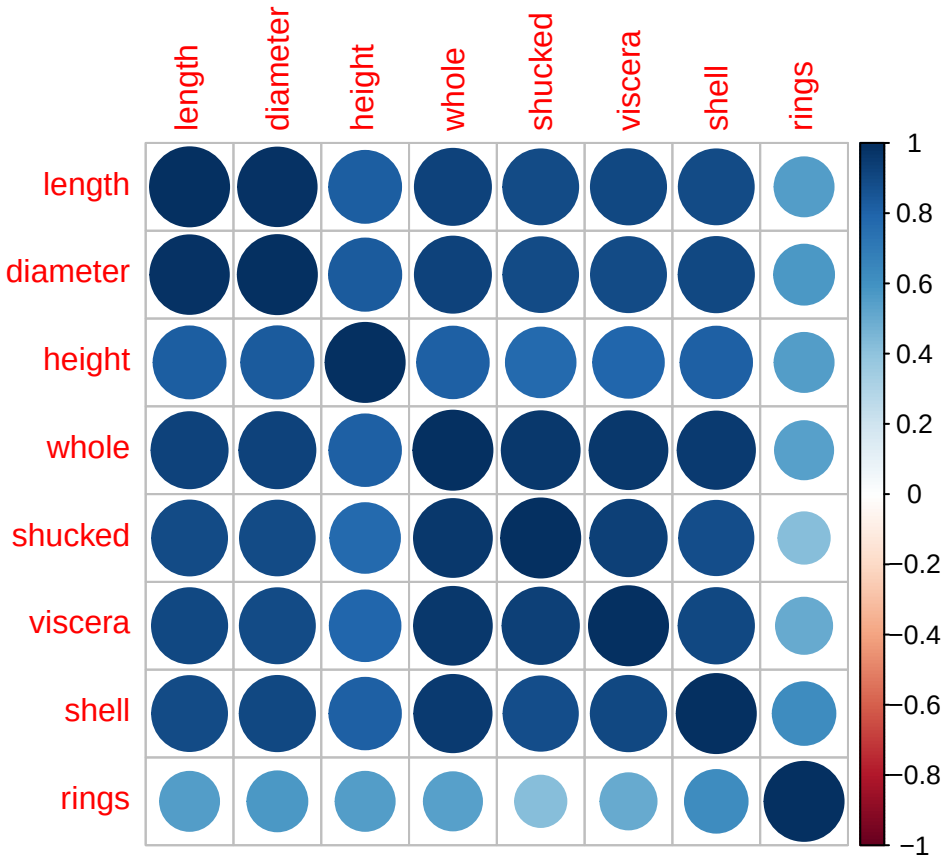
Box plots


boxplot
length
diameter
height
whole
shucked
viscera
shell
rings

Pairs plot


sex_num
length
diameter
height
whole
shucked
viscera
shell
rings

Correlation graphic



Analysis

Regression and classification analyses are done using the following code:

```
fits <- lapply(models, function(model) {
  print(model)
  set.seed(seed)
  train(
    rings ~ .,
    data = train_set_r,
    method = model,
    metric = metric,
    trControl = control
  )
})
```

There is some variation in the code depending upon whether regression or classification.

Regression Analysis

The result from running regression analysis is:

method	RMSE
rf	2.120327
gbm	2.148518
knn	2.183692
lm	2.205379
glm	2.205379
svmLinear	2.236187
treebag	2.244404
gamLoess	2.346836
rpart	2.612525

"rf" is the best model based on a minimum RMSE of 2.12033.

Using the "rf" model with test data, the accuracy is: 0.26253.

Classification Analysis

The result of running classification analysis is:

method	Accuracy
svmLinear	0.7380862
multinom	0.7360463
svmRadial	0.7333803
lda	0.7321401
knn	0.7213184
rf	0.7186519
qda	0.6702152
gamLoess	0.6257860
naive_bayes	0.5663580

Based on comparing accuracy for all the models, using training data, "svmLinear" is the best model with an accuracy of 0.73809.

Using the "svmLinear" model with the test data, the accuracy is: 0.75179

Results

Regression and classification analyses are done for multiple models. The best result is obtained with classification using the "svmLinear" model by segregating rings into age groups. Accuracy using test data is: 0.75179.

Conclusion

Given the many variables, such as availability of food, water conditions, and weather patterns, which are outside the scope of this dataset, predicting rings (i.e. age) is difficult. Grouping age and analyzing by classification, we can get a fair approximation.

Additional work can be done on tuning the models, exploring other methods, and/or including additional data, such as weather, water conditions, location, and food availability.

Citations

The Editors of Encyclopaedia Britannica. (2020, April 23). *Abalone*. Encyclopedia Britanica. https://www.britannica.com/animal/abalone