Abalone Age Prediction

Jose Luis Estrada, Nava Roohi, and Ashutosh Singh

6/20/2021

Traditionally, the process to predict the age of the abalone is by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. The abalone dataset includes numeric attributes with different type of measurements with the goal to predict the age of an abalone more efficiently.

```
abalone <- read.csv('abalone.csv', header = TRUE)
abalone$Age <- abalone$Rings+1.5
abalone <- subset(abalone, select = -c(Rings))
```

The table has 4,177 observations and 9 columns. The attribute Rings was replaced by the variable Age since each rings is equivalent to the number of rings plus 1.5 (Hossain). This will be helpful at a later stage of the project when the dataset is split into training and test set. As a reminder, the goal of this project is to calculate the age of the abalone, so the dependent variable will be Age.

```
summary(abalone)
```

```
##      Sex                Length         Diameter          Height
##  Length:4177        Min.   :0.075   Min.   :0.0550   Min.   :0.0000
##  Class :character   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150
##  Mode  :character   Median :0.545   Median :0.4250   Median :0.1400
##                     Mean   :0.524   Mean   :0.4079   Mean   :0.1395
##                     3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650
##                     Max.   :0.815   Max.   :0.6500   Max.   :1.1300
##   Whole.weight    Shucked.weight   Viscera.weight    Shell.weight
##  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
##       Age
##  Min.   : 2.50
##  1st Qu.: 9.50
##  Median :10.50
##  Mean   :11.43
##  3rd Qu.:12.50
##  Max.   :30.50
```

Furthermore, the attributes in the abalone dataset are numeric in its majority, and Sex is the only categorical data (binary). The dataset is not missing any values, so no data cleaning is needed.

```r
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.weight  : num  0.514 0.226 0.677 0.516 0.205 ...
##  $ Shucked.weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
##  $ Viscera.weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
##  $ Shell.weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
##  $ Age           : num  16.5 8.5 10.5 11.5 8.5 9.5 21.5 17.5 10.5 20.5 ...
```

```r
sapply(abalone, function(x) sum(is.na(abalone)))
```

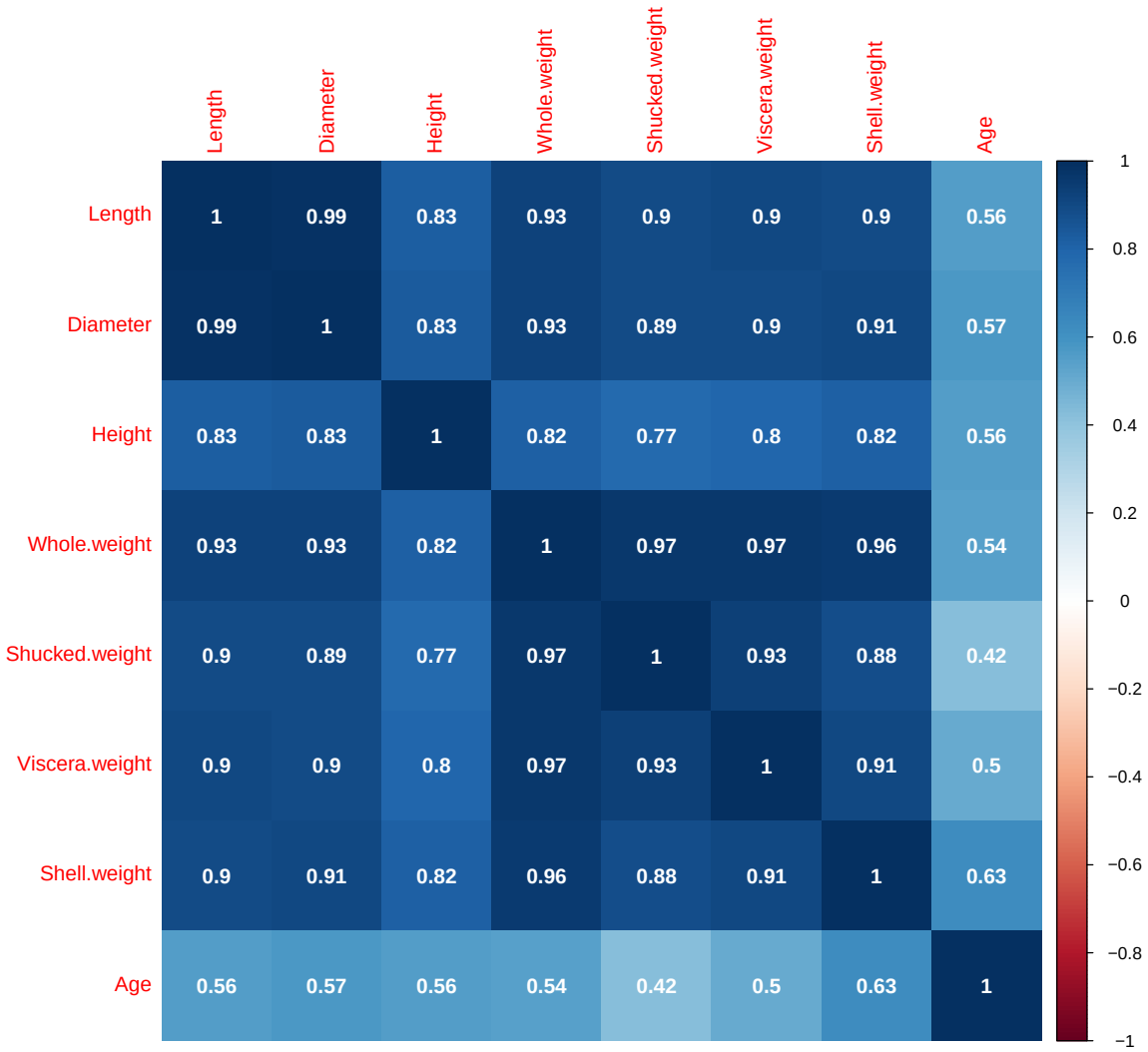
```
##            Sex         Length       Diameter         Height   Whole.weight
##              0              0              0              0              0
## Shucked.weight Viscera.weight   Shell.weight            Age
##              0              0              0              0
```

The correlation plot shows high correlations (between 0.75 and 0.99) within the independent variables, but a medium direct correlation with the target value (between 0.4 and .65)

```r
library(corrplot)
```

```
## corrplot 0.88 loaded
```

```r
corrplot(cor(abalone[c(2:9)]), method = "shade", addCoef.col = "white")
```


Length
Diameter
Height
Whole.weight
Shucked.weight
Viscera.weight
Shell.weight
Age
Length 1 0.99 0.83 0.93 0.9 0.9 0.9 0.56
Diameter 0.99 1 0.83 0.93 0.89 0.9 0.91 0.57
Height 0.83 0.83 1 0.82 0.77 0.8 0.82 0.56
Whole.weight 0.93 0.93 0.82 1 0.97 0.97 0.96 0.54
Shucked.weight 0.9 0.89 0.77 0.97 1 0.93 0.88 0.42
Viscera.weight 0.9 0.9 0.8 0.97 0.93 1 0.91 0.5
Shell.weight 0.9 0.91 0.82 0.96 0.88 0.91 1 0.63
Age 0.56 0.57 0.56 0.54 0.42 0.5 0.63 1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1