Abalone HarvardX Capstone

Ana Hristova

18/06/2020

1. Introduction

Abalone are shellfish from the Haliotidae family known for their pearlescent inside shell living in seas and oceans all around the world. The age of an abalone is determined by opening its shell, cleaning it and then counting the rings, which is a repetitive and time-consuming task.

The aim of this project is to examine data driven models that help estimate the age of an abalone by using its measurements which are more easily obtainable.

2. Data

The full abalone dataset contains 4177 observations of abalone sex and measurements. It is originally provided in the University of Irvine data library and can be accessed via this link: https://archive.ics.uci.edu/ml/datasets/abalone

The data contains no NAs. For the purposes of creating a verifiable predictive model, the data is split into "abalone" and "validation" sets with a 90% and 10% split respectively. The validation set is only used after choosing a best fitting model to validate its predictions.

3. Methodology

As the aim of this project is to create a predictive model that predicts the number of rings of abalones, simulating a real modelling environment, only the post-split abalone data set was used for the exploratory analysis of the data.

The exploratory analysis of the data includes deep dive into each variable, to understand any dependancies and data types included.

Following the exploratory analysis, the abalone dataset was split into train and test set to allow for examining different models and variables and to arrive to a model that would deliver the lowest RMSE. RMSE was chosen as a success measurement as the predicted variable (number of rings) is non-categorical.

The predictive models explored are:

1. Principal component regression
2. Partial least squares
3. k Nearest Neigbors

The 90%/10% split used for splitting the original abalone set into abalone and validation sets and for splitting the resulting abalone set into test and train sets was chosen to provide a large enough number of observations to train the model but also to retain enough observations to produce a validation that didn't simply happen by chance.

4. Exploratory data analysis

4.1. Exploring the data set

The abalone dataset contains 9 variables and 3758 observations. The variables included are: sex, length, diameter, height, whole, shucked, viscera, shell, rings

sex	length	diameter	height	whole	shucked	viscera	shell	rings
M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7
I	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.120	8

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3758 obs. of  9 variables:
##  $ sex     : Factor w/ 3 levels "F","I","M": 3 3 1 3 2 2 1 1 3 1 ...
##  $ 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   : num  15 7 9 10 7 8 20 16 9 19 ...
```

We have two types of variables:

- sex: factor with three levels “M” (male), “F” (female) and “I” (infant)
- the rest of the variables are all numeric - rings is discreet and the measurements are continuous

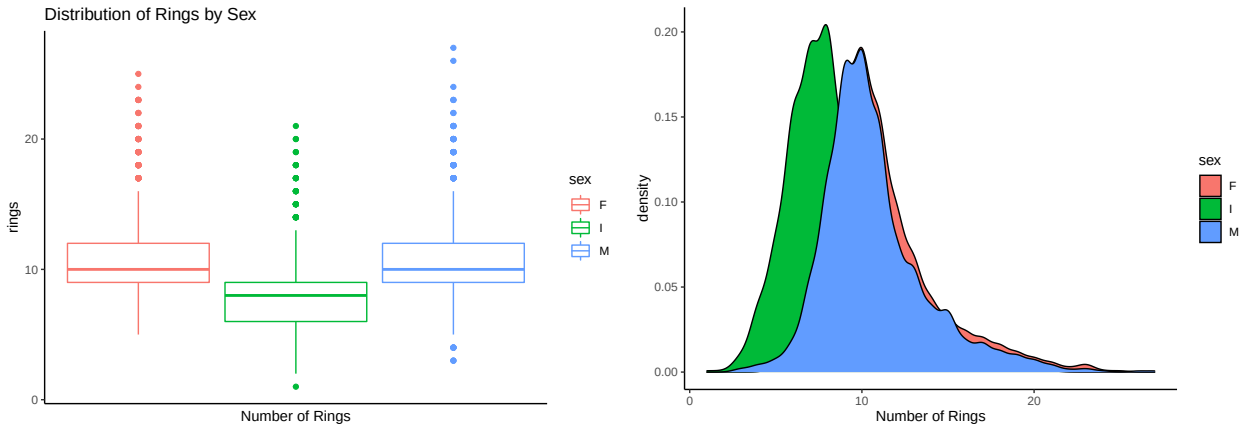
4.2. Exploring the sex variable

It appears we have a evenly distributed number of infants, males and females in our dataset:

sex	count
F	1178
I	1202
M	1378

Let us have a look at the average number of rings for each sex and their distribution:

sex	avg_rings
F	11.136672
I	7.874376
M	10.708273



From the graphs we see that while infants clearly tend to have fewer rings than adult abalones, adult male and female abalones have a very similar distribution and their sex is not a clear predictor of their number of rings and therefore also age.

We can use this insight when developing our models by turning sex to binary adult = 1 and infant = 0 for models that cannot handle non-numerical data such as linear regression.

4.3. Exploring the numeric variables

From the exploration of the dataset above, we know we have 7 numeric explanatory variables: * Hieght * Length * Diameter * Whole weight * Viscera weight * Shucked weight * Shell weight

Let's find out more about them:

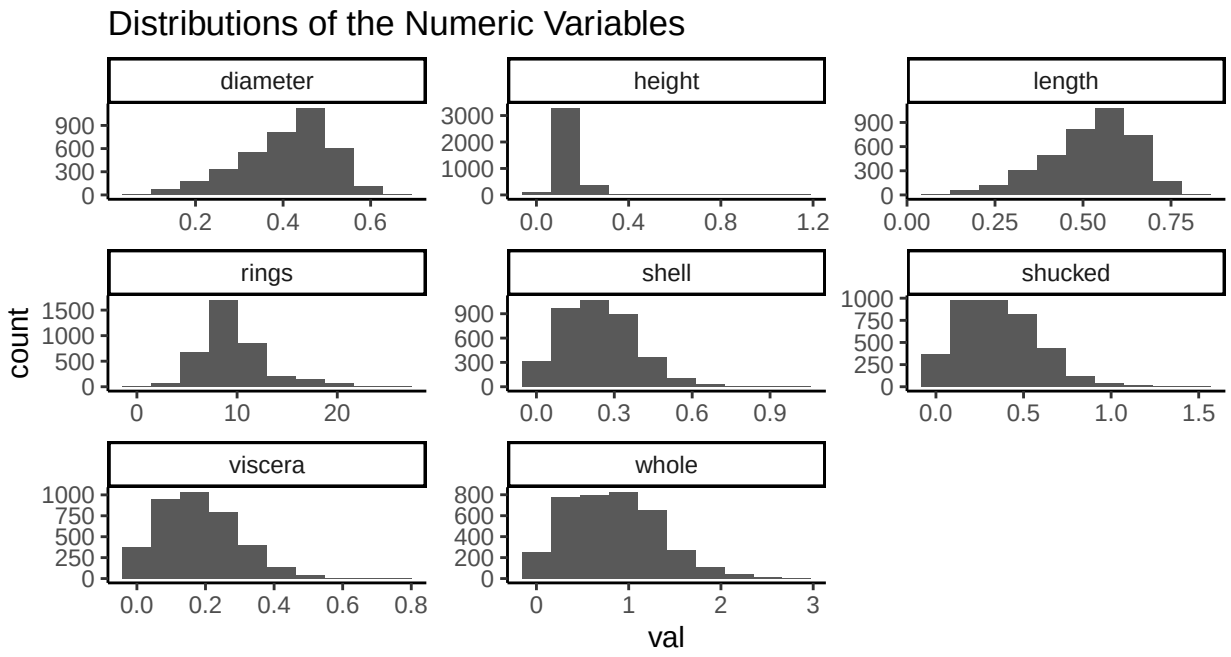
length	diameter	height	whole	shucked	viscera	shell
Min. :0.075	Min. :0.0550	Min. :0.0000	Min. :0.0020	Min. :0.0010	Min. :0.00050	Min. :0.0015
1st Qu.:0.450	1st Qu.:0.3500	1st Qu.:0.1150	1st Qu.:0.4405	1st Qu.:0.1860	1st Qu.:0.09312	1st Qu.:0.130
Median :0.545	Median :0.4250	Median :0.1400	Median :0.8035	Median :0.3372	Median :0.17100	Median :0.235
Mean :0.524	Mean :0.4078	Mean :0.1394	Mean :0.8293	Mean :0.3596	Mean :0.18088	Mean :0.2389
3rd Qu.:0.615	3rd Qu.:0.4800	3rd Qu.:0.1650	3rd Qu.:1.1540	3rd Qu.:0.5040	3rd Qu.:0.25400	3rd Qu.:0.329
Max. :0.815	Max. :0.6500	Max. :1.1300	Max. :2.8255	Max. :1.4880	Max. :0.76000	Max. :1.0050

Itr appears we have some abalones in the dataset with height = 0. We need to take a closer look to determine wether to keep these entries or discard them.

sex	length	diameter	height	whole	shucked	viscera	shell	rings
I	0.430	0.34	0	0.428	0.2065	0.0860	0.1150	8
I	0.315	0.23	0	0.134	0.0575	0.0285	0.3505	6

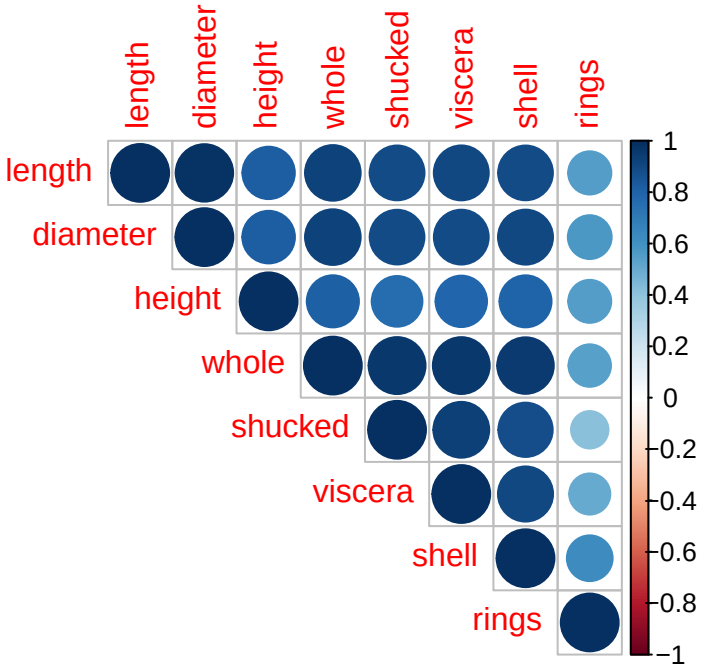
Whole height is 0 for two abalones, all their other measurements are present, so even if this is an entry error we cannot discard them and will retain them in the dataset.

Now we will look at the distributions of the numeric variable sand the rings to see if there are similar patterns:



The plot above shows that rings and the weight measurements have similar, right-skewed distribution. This similarity in patterns indicates they will likely be better predictors then the size measurements.

Next we will examine the correlation between the numeric variables and the rings:



All numeric variables are moderately positively correlated with the number of rings of the abalones in the dataset. However, we notice that the correlation between the explanatory variables is a lot stronger, almost perfect in some cases. This information, while not surprising as the larger the abalone, the more it weights, presents the challeng of multicollinearity in our data and helps us determine which models will be useful and which will result in overtraining the model.

Due to this multicollinearity between the explanatory variables, linear regression will result in an overtrained

model which will produce worse results the more abalones we include in out validation set. To avoid this, we will use methods that are not affected by multicollinearity, such as principal component regression, partial least squares and k nearest neighbors.

5. Choosing the best predictive model

In this section we will take a look at 3 predictive models to choose the best one amongst them, which will yield the lowest RMSE value.

For this purpose, the abalone dataset explored above is split into a train and test set, which will allow for the model to be tested and fitted before applying the best one to the validation set.

5.1. Principal Component Regression

PCR works by applying Principal Component Analysis to the data and summarising the original predictor variables into new variables called Principal Components, which are a linear combination of the original data. Then these Principal Components (PCs) are used for building a linear regression model. The number of PCs included in the model is determined by cross-validation.

```
set.seed(1, sample.kind = "Rounding")
```

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
train_pc<- train(
  rings~.-sex, data = train, method = "pcr",
  scale = TRUE,
  trControl = trainControl("cv", number = 7),
  tuneLength = 7
)

train_pc$bestTune
```

```
##   ncomp
## 6     6
```

```
fit_pc<- train_pc$finalModel
pred_pc<- predict(fit_pc, test)
pc_reg<-RMSE(test$rings, pred_pc)
```

method	RMSE
Principal component regression	2.350341

RMSE of 2.35 is not too bad but PCR gives no guarantee that the selected PCs are associated with the outcome. Next we will look at Partial Least Squares regression, which takes this into account.

5.2. Partial Least Squares regression

Partial Least Squares regression is an alternative to PCR which also identifies new PCs that summarise the original predictors but also it makes sure they are related to the outcome. Again we will use cross-validation to determine the best number of PCs.

```
# Partial least squares
set.seed(1, sample.kind = "Rounding")
train_pls <- train(
  rings~.-sex, data = train, method = "pls",
```

```
  scale = TRUE,
  trControl = trainControl("cv", number = 7),
  tuneLength = 7
)

train_pls$bestTune
```

```
##   ncomp
## 6     6
```

```
fit_pls<- train_pls$finalModel
pred_pls<- predict(fit_pls, test)
pls<-RMSE(test$rings, pred_pls)

rmse_results<- bind_rows(rmse_results, data_frame(method = "Partial least squares",
                                                  RMSE = pls))

knitr::kable(rmse_results)
```

method	RMSE
Principal component regression	2.350341
Partial least squares	2.251076

We have an improvement of 0.1 on our RMSE, which is not bad. Let's try one more model to see if we can do better.

5.3. k Nearest Neigbors

k Nearest Neighbors is quite different from the other two models we used above, as instead of using an altered linear model it uses similarity measure to predict the dependant variable, in our case the number of rings and therefore the age of an abalone.

Let's have a look at how this model performs:

```
fit_knn<- train(rings~., method = "knn", data=train)
pred_knn<- predict(fit_knn, test)
knn_model<-RMSE(test$rings,pred_knn)

rmse_results<- bind_rows(rmse_results, data_frame(method = "k Nearest Neigbors",
                                                  RMSE = knn_model))

knitr::kable(rmse_results)
```

method	RMSE
Principal component regression	2.350341
Partial least squares	2.251076
k Nearest Neigbors	2.090100

The RMSE is now 2.09, whooping 0.26 lower than PCR and 0.16 lower than PLS. Therefore, this is the model we choose as the best one and we will test it on the validation data.

6. Results

The k Nearest Neighbors model tested above showed the most promising predictions, with RMSE of 2.09. Therefore, it is the one we choose to keep for predicting the number of rings of an abalone given its sex and measures.

```
# Predicting using the kNN model: ----

fit_knn_final<- train(rings~., method = "knn", data=abalone)
pred_knn_final<- predict(fit_knn_final, validation)

# RMSE result ----
result<-RMSE(validation$rings, pred_knn_final)
```

It's RMSE of 2.1537595 is slightly worse than the one obtained with the test and train sets, but it still performs better than the other two models explored.

7. Conclusion

This report took us through the abalone dataset used for predicting the age of an abalone based on its sex and measurements.

The entire set was split 90/10 to achieve abalone and validation sets, used for modelling the data and validating the final model respectively.

The set contained 4177 observations of 9 variables wich were explored in detail in the Exploratory Analysis section of this report.

The predictive model that best fit the data and provided the lowest RMSE amongst the models explored was k Nearest Neigbors.The results of each of the models explored are summarised in the table below:

method	RMSE
Principal component regression	2.350341
Partial least squares	2.251076
k Nearest Neigbors	2.090100

8. Limitations and further work

Only 3 models were explored in predicting the number of rings in the validation set. The small size of the data set and the high multicollinearirty of the explanatory numeric variables made modelling the data easy to overtrain.

Working with a larger dataset with more entries and data on the habitats and diets of the abalones will be necessary to develop a more accurate model.