Abalone Intro Exercise

Evan Hope

2024-03-15

Introduction

The purpose of this exercise is to display my introductory knowledge of reading in a data set, appending a variable to the set, setting up a workflow, creating a recipe, splitting data into training and test sets, and using the recipe and training set to help create predictions of the age of an abalone.

Below is a walk through of my work.

Package Installation

Before I begin, I will now download the packages needed to conduct data manipulation and modeling.

```
#install.packages("tidyverse")
```

```
#install.packages("tidymodels")
```

```
#install.packages("ISLR")
```

```
#install.packages("dplyr")
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.1.1 --
## v broom        1.0.5      v rsample      1.2.0
```

```
## v dials        1.2.1      v tune         1.1.2
## v infer        1.0.6      v workflows    1.1.4
## v modeldata    1.3.0      v workflowsets 1.0.1
## v parsnip      1.2.0      v yardstick    1.3.0
## v recipes      1.0.10
## -- Conflicts ---------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(ISLR)
library(dplyr)
```

Reading in the Data

Now that the necessary packages are ready to go, I will now read in the abalone data set.

```
abalone_data = read.csv("C:/Users/Ordai/OneDrive/Desktop/School/Personal Projects/Abalone Intro Exercis
```

I will slightly modify the data set by adding an "age" variable.

```
abalone_data2 <- mutate(abalone_data, age = rings + 1.5)
# The age (in years) of an abalone can be quickly estimated by adding 1.5 to the number of rings found

glimpse(abalone_data2)
```

```
## Rows: 4,177
## Columns: 10
## $ type           <chr> "M", "M", "F", "M", "I", "I", "F", "F", "M", "F", "F", ~
## $ longest_shell  <dbl> 0.455, 0.350, 0.530, 0.440, 0.330, 0.425, 0.530, 0.545,~
## $ diameter       <dbl> 0.365, 0.265, 0.420, 0.365, 0.255, 0.300, 0.415, 0.425,~
## $ height         <dbl> 0.095, 0.090, 0.135, 0.125, 0.080, 0.095, 0.150, 0.125,~
## $ whole_weight   <dbl> 0.5140, 0.2255, 0.6770, 0.5160, 0.2050, 0.3515, 0.7775,~
## $ shucked_weight <dbl> 0.2245, 0.0995, 0.2565, 0.2155, 0.0895, 0.1410, 0.2370,~
## $ viscera_weight <dbl> 0.1010, 0.0485, 0.1415, 0.1140, 0.0395, 0.0775, 0.1415,~
## $ shell_weight   <dbl> 0.150, 0.070, 0.210, 0.155, 0.055, 0.120, 0.330, 0.260,~
## $ rings          <int> 15, 7, 9, 10, 7, 8, 20, 16, 9, 19, 14, 10, 11, 10, 10, ~
## $ age            <dbl> 16.5, 8.5, 10.5, 11.5, 8.5, 9.5, 21.5, 17.5, 10.5, 20.5~
```

Note that my end goal of this exercise is to predict the age of an abalone by creating a recipe that uses a combination of variables EXCLUDING the "rings" variable. The "rings" variable as well as the formula used above (rings+1.5=age) will only be used as a reference to see how close my recipe comes to predicting the actual age (rings + 1.5). I will expand a little more on this shortly.

Data Splitting

I will now set a seed and split the data.

```
set.seed(909)
# Setting a seed in order to get reproducible results.

abalone_split <- initial_split(abalone_data2, prop = 0.80, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Creating a Recipe

The reason why I am excluding the rings variable here is because I have asserted that there is a direct relationship between the number of rings and age. If I were to resume my analysis with the "rings" variable included, my prediction model would tell me that the rings variable is the only variable that matters when predicting the age. But for the sake of this exercise, I want to show that I can create a model that can predict the age of an abalone without using the rings variable.

```
# First a simple recipe to reference from using the training set.

simple_abalone_recipe <- recipe(age ~ ., data = abalone_train)
```

```
# Dummy coding the categorical variables while also centering and scaling the predictors...

abalone_recipe <- recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_weight
  step_dummy(all_nominal_predictors()) %>%
  step_normalize(all_predictors())
```

Creating a LInear Regression Model & Workflow

```
#Setting up a linear regression model

lm_model <- linear_reg() %>%
  set_engine("lm")
```

```
#Creating the workflow and adding our recipe

lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)
```

```
#Fitting the linear model to the abalone training data
lm_fit <- fit(lm_wflow, abalone_train)
```

Results of the predictor variables

```
lm_fit %>%
  # This returns the parsnip object:
  extract_fit_parsnip() %>%
  # Now tidy the linear model object:
  tidy()
```

```
## # A tibble: 10 x 5
##    term           estimate std.error statistic  p.value
##    <chr>             <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)     11.4       0.0379    302.   0
##  2 longest_shell   -0.244     0.244      -1.00 3.17e- 1
##  3 diameter         1.33      0.248       5.35 9.56e- 8
##  4 height           0.425     0.0696      6.11 1.11e- 9
##  5 whole_weight     4.16      0.408      10.2  4.55e-24
##  6 shucked_weight  -4.21      0.205     -20.6  1.66e-88
##  7 viscera_weight  -1.16      0.160      -7.25 5.31e-13
##  8 shell_weight     1.31      0.180       7.25 5.30e-13
##  9 type_I          -0.376     0.0536     -7.02 2.77e-12
## 10 type_M           0.0556    0.0449      1.24 2.15e- 1
```

Notice the p-values for each variable. Every variable EXCEPT the "longest shell" length and "type" for males has a p-value less than 0.05. Those that have a p-value less than 0.05 are known to be statistically significant.

Predicting the age of an abalone using the new model!

I will now see what age is predicted when I create a data frame (abalone) that is a female ("I") whose longest shell is .50 cm, has a diameter of .10 cm, a height of .30 cm, weighing 4 kg, etc.

```
female_abalone_pred1 <- data.frame(type = 'I', longest_shell =.50, diameter = 0.10, height = 0.30, whol

abalone_train_pred = predict(lm_fit, new_data = female_abalone_pred1)

abalone_train_pred
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  10.9
```

As we can see, my model predicted that the age, given the information above, would be about 10.9 years.

I will now come up with predictions for all of the observations within the test set.

```
library(yardstick)
abalone_test_pred = predict(lm_fit, new_data = abalone_test)

abalone_test_pred
```

```
## # A tibble: 837 x 1
##    .pred
##    <dbl>
##  1 12.5
##  2 11.2
##  3 12.3
##  4 10.3
##  5 12.0
##  6  8.89
##  7 10.2
##  8 10.7
##  9 10.3
## 10 10.4
## # i 827 more rows
```

Model Performance and Metrics

I will now create one last data set that combines the predictions that I have just computed above with the original test set in order to measure the accuracy of my model.

```
abalone_test_v2 <- mutate(abalone_test, abalone_test_pred)
#combining the predictions with original test set

abalone_metrics <- metric_set(rmse, rsq, mae)
#calculating the root mean squared error, R-squared, and mean absolute error

abalone_metrics(abalone_test_v2, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.22
## 2 rsq     standard       0.516
## 3 mae     standard       1.60
```

As we can see from the metrics calculated above, the linear regression model used to predict the age of an abalone based on several factors (excluding the # of rings) performed well.