Individual Final Report

Ziyu Huang (G28153536)

1. Introduction

The project aims at predicting the age of abalone by using a series of machine learning method to deal with data sets. The outline of the shared work includes applying linear regression method to the data processed by my groupmate to making prediction. And then, getting the result and compare the result with that of other models my groupmates make.

2. Description of my individual work

2.1 Description

My work is to draw the plot including the histogram plot and pair plot to analyze the correlation relationship between features, and to apply normalization and standard scaler to deal with data set to see the different influence of applying model, and to use linear regression model to predict the age of the abalone. In addition, comparing the result of different model my groupmates make, and drawing the conclusion.

2.2 Background information

2.2.1 Linear regression

Given a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressor x is linear. The error variable ε is an unobserved random variable that add "noise" to the linear relationship between y and xi. The model is shown as below:

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i$, i = 1, …, n,

Where T denotes the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β.

It can be also written in matrix notation as

$y = X\beta + \varepsilon$,

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_1 \end{pmatrix}$$

$$X = \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

2.2.2 Evaluation index

Root Mean Square Error (RMSE)

Root Mean Square Error is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are, and RMSE is a measure of how spread out these residuals are. Normally speaking, the smaller RMSE, the better the model.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}}$$

R-squared

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The closer R-squared to 1, the better performance the model has.

$$R - squared = \frac{Explained\ variation}{Total\ variation}$$

3. My portion on the project

First, I draw the count plot of the data set to see how many data each sex includes, and draw the pair plot to see the distribution of the single variable and the relationship between variables.

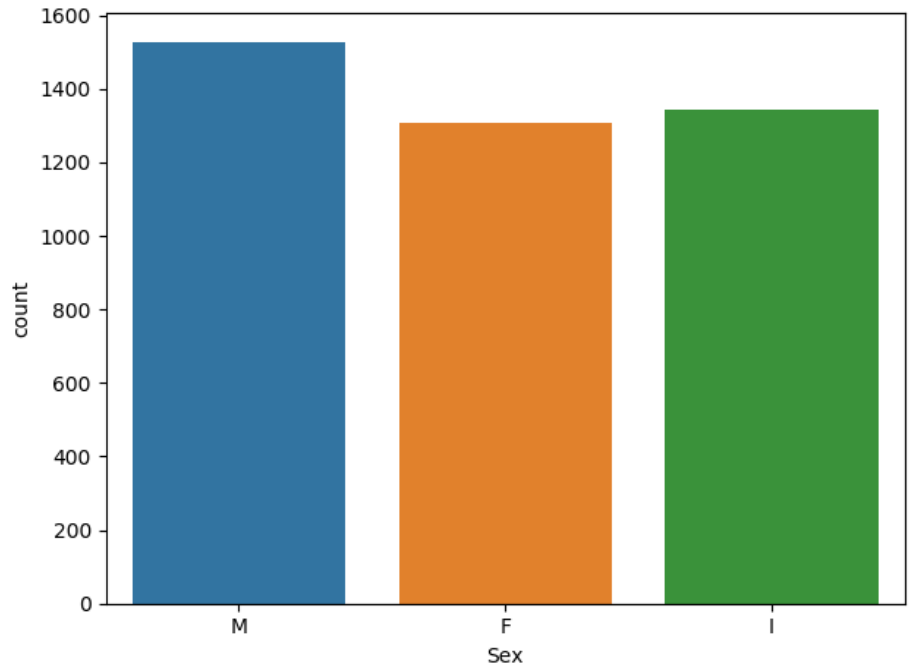
Next, I split the data set to training data and test, and make standardization and normalization to the data set separately, and compare the effect of these two method to the result we finally get.

Then, I apply linear regression model to the data set that we finally have processed. I calculate the score, Mean Squared Error and R-squared of this model to help us analyze the result, and draw the plot to view the result intuitively.

In the last, I compare the score of the linear regression model to that of my groupmates' logistic regression and multiple layer perceptron to get the appropriate model to predict the age of abalone.

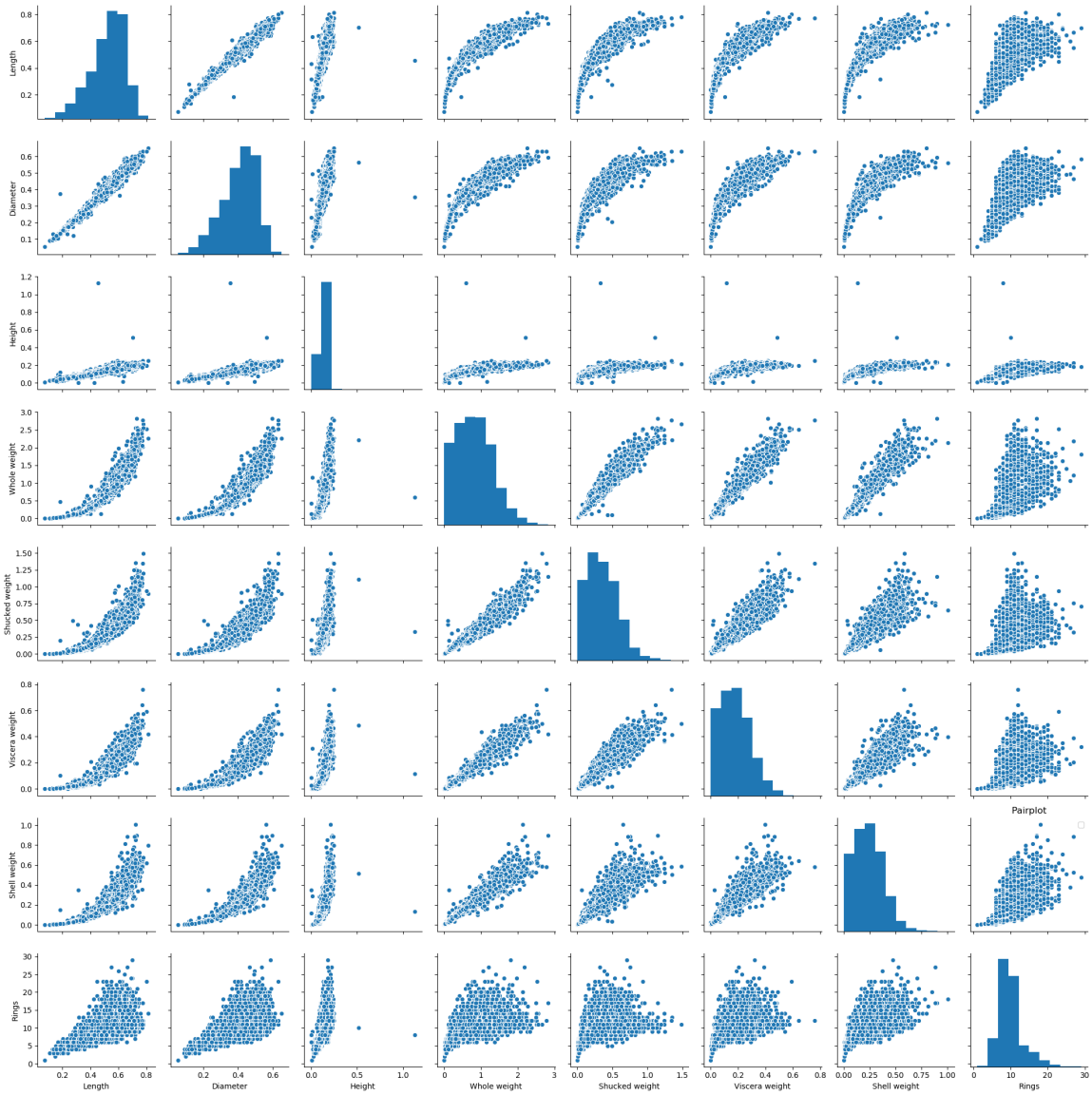
4. Results

4.1 Count plot



From the count plot we get above, the data separate roughly equally to the male, female and infant.

4.2 Pair plot



The pair plot shows the distribution of each variable and the relationship between variables. As we can see from the graph, there are positive relationships between variables. In addition, there are outliers in some graphs, so addition data preprocessing is needed to be done.

4.3 Data preprocessing

4.3.1 Standardization

```
StandardScaler_train [[ 0.21935428  1.16602996]
 [ 1.09013478 -0.58906117]
 [-1.35726558 -0.66875823]
 ...
 [ 1.37548018  0.86538308]
 [ 0.19219773  1.02758631]
 [ 0.91266412 -1.49236626]]
StandardScaler_test [[ 0.17114712  0.92426609]
 [ 1.2477301   0.21140746]
 [ 1.38917454  0.93870123]
 ...
 [ 0.9406967  -1.34881909]
 [ 1.03244846 -0.88226441]
 [ 0.26098533  1.38059099]]
```

```
print result of lr:
score:  0.3019867549668874
mse:  2.6494631577077326
rms:  1.6277171614588735
r^2:  0.5483131246344208
```

The result after we standardize the abalone data and using linear regression method to predict the age is shown as above. The score is 0.302, RMSE is 1.628, and R-squared is 0.548.

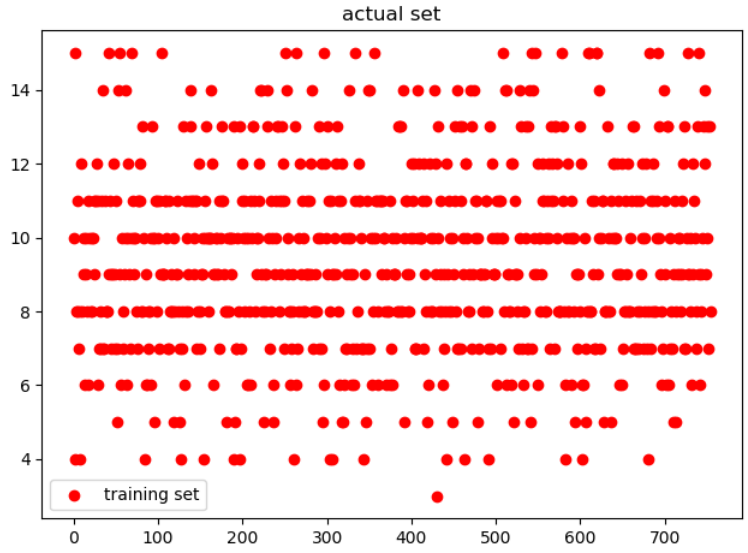
4.3.2 Normalization

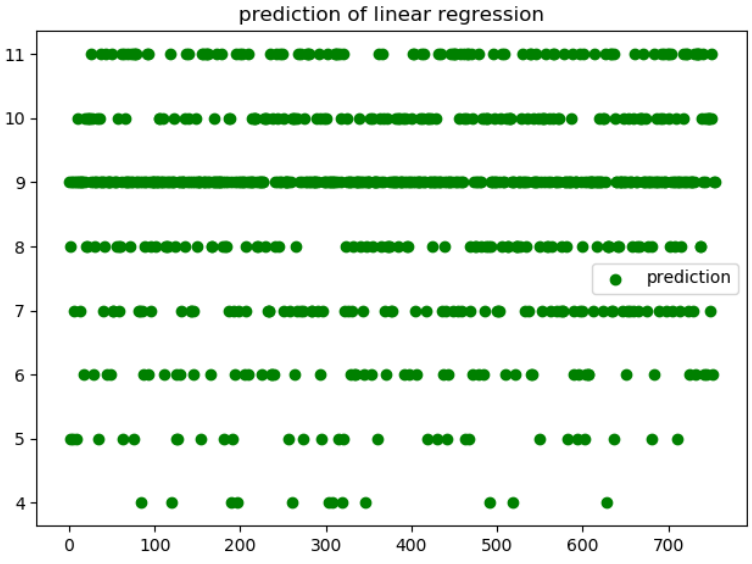
```
Normalization_train [[ 0.96184863 -0.27358219]
 [-0.98740039  0.15824183]
 [ 0.86289115 -0.50538982]
 ...
 [-0.88176728  0.4716847 ]
 [-0.20964179 -0.97777826]
 [-0.87594496 -0.48241105]]
Normalization_test [[-0.27069691 -0.96266463]
 [-0.88154386 -0.47210214]
 [-0.95699522 -0.29010369]
 ...
 [ 0.17627538  0.98434089]
 [ 0.18843199  0.98208624]
 [ 0.88848721 -0.45890138]]
```

```
print result of lr:
score:  0.2635761589403974
mse:  3.948484322925815
rms:  1.9870793448993966
r^2:  0.27137012279206973
```

The result after we normalize the abalone data and using linear regression method to predict the age is shown as above. The score is 0.263, RMSE is 1.987, and R-squared is 0.271. As we can see, the score of standardization is bigger than that of normalization, RMSE of the former one is smaller than the later one. In addition, the R-squared of the first one is bigger than the second one, so the standardization is better than normalization. As a result, we use standardization to preprocess the data.

4.4 Plot of linear regression





The plot of the original data set and result of linear regression prediction are shown as above. As we can see in the graph, the data match not as good as we expected.

5. Summary and conclusion

5.1 Analyze linear regression model

From the results we get above, we can conclude that standardization is more proper to linear regression than normalization, and the effect of linear regression predicting the age of abalone is not that good since R-squared is less than 0.7. As a result, we may conclude that linear regression is not suitable for fitting this abalone data set, and a better model is needed to predict the age of the abalone.
However, we need to compare the linear regression model to the model that my groupmates make, random forest regression, multiple layer perceptron, and logistic regression.

5.2 Compare the different models

In our report, regression algorithms and classification algorithms are used to predict the target. For the regression algorithm, we use linear regression and random forest. Root of Mean Squared Error (RMSE) and R-Squared are used to compare the model. As for the classification algorithm, multiple layer perceptron and logistic regression, accuracy score is used to compare the models. The detailed information are shown as follows:

Regression algorithms		
Model	RMSE	R-Squared
Linear regression	1.648	0.548
Random Forest	2.104	0.535

From the table above, R-Squared of random forest is bigger than that of linear regression, it turns out that the effect of random forest is better. However, RMSE of random forest is bigger than that of linear regression. Since we expect that the better model has smaller RMSE, another model is needed to predict the target.

Classification algorithms			
Model	RMSE	R-Squared	Accuracy score
Multiple layer perceptron	0.55	-0.39	0.70
Logistic regression	0.46	-0.004	0.72

As we can see, accuracy score of multiple layer perceptron is 0.70, and accuracy score of logistic regression is 0.72, so logistic regression has the better performance than another. However, it is not significantly superior to the other one, so further analysis is needed to be done in the future.

5.3 Limitation and Improvement

The performance of our model is not as good as what we have expected, probably due to the imbalanced data set. The solution to this problem is reducing the number of class into binary class. The model can predict general range of age of abalone. Other solutions involve increasing more instance and features by provide some feature engineering.

6. Calculate the percentage of the code I from on the internet.

I use 18 lines of code from the internet and modify 10 lines, and I add another 24 lines of my own code, so the percentage is 19%.

7. References

[1] "Abalone Dataset." *Kaggle*, 19 July 2018, www.kaggle.com/rodolfomendes/abalone-dataset.

[2] Ragnisah. "EDA- Abalone Age Prediction." *Kaggle*, 28 Aug. 2019, www.kaggle.com/ragnisah/eda-abalone-age-prediction.

[3] Anthonypino. "Price Analysis and Linear Regression." *Kaggle*, 16 Aug. 2017, www.kaggle.com/anthonypino/price-analysis-and-linear-regression.