We employed 4 method. They are NaïveBayes, Logistic Regression, Decision Tree, K-Nearest Neighbor.

Training data is 60%, Validation data is 20% and test data is 20% from data.

I ,

NaïveBayes:

First, we deleted highly correlated variables. We keep height, Shucked weight, Rings.

We use 60% of data as training data, and 20 % of data is validation data.

F M

0.4480712 0.5519288

Conditional probabilities:

data.height

trainData[, 1] [,1] [,2]

F 0.1576159 0.02895067

M 0.1524570 0.03482404

data.Sweight

trainData[, 1] [,1] [,2]

F 0.4497828 0.1996874

M 0.4387710 0.2249638

data.Rings

trainData[, 1] [,1] [,2]

F 11.19735 3.115716

M 10.67849 2.853976

P(Female):0.448, P(Male): 0.552

predicted F M

F 59 63

M 205 244

The power rate is 0.53

II

Logistic regression:

There is no statistical assumption under Logistic regression, so, we did not change the variables.

Model:

Coefficients:

(Intercept) length Diameter height Wweight Sweight

2.79073 -2.38368 -3.82736 -0.91203 0.38191 2.24023

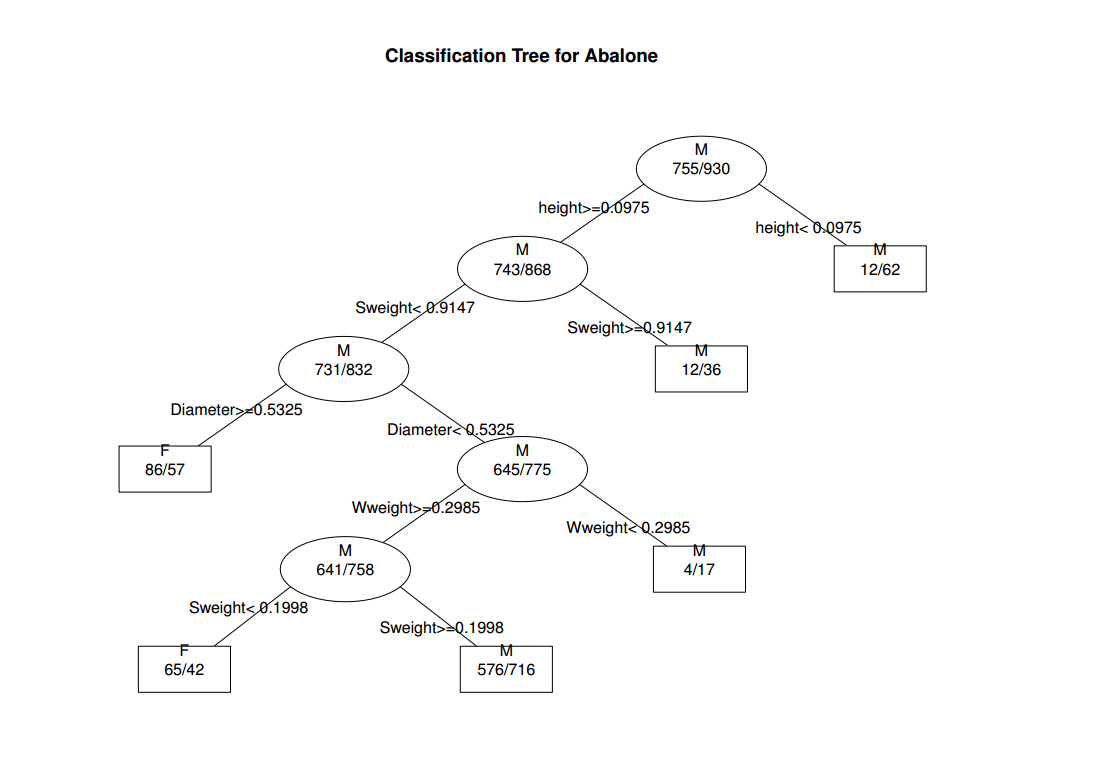
Vweight Shellweight Rings

-1.61620 -0.67096 -0.01648

The power rate is 0.553. The best cutoff point is 0.02.

III

Decision Tree.



We use the Traing data to plot the tree. The package called rpart in R plotted for us. It generated the best pruned tree.

When we put the Validation data into this tree. The power rate is 0.5516.

VI

K-Nearest Neighbor. (K is 7)

We used Euclidean distance to calculate. We didn’t delete variable. The power rate is 0.512.

Result:

|  |  |  |  |
| --- | --- | --- | --- |
| Naive | tree | Logistic | K-NN |
| 0.5288967 | 0.5516637 | 0.553 | 0.5288967 |

The table conclude all the results from 4 methods. Logistic regression gave us highest rate, and K-nearest neighbor gave the worst result.

Conclusion:

How to choose the best model?

We use cross validation to select the best model.

Cross-validation is a [model validation](http://en.wikipedia.org/wiki/Model_validation) technique for assessing how the results of a [statistical](http://en.wikipedia.org/wiki/Statistics) analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how [accurately](http://en.wikipedia.org/wiki/Accuracy) a predictive model will perform in practice.

We decided to test 10 combinations of training data set and validation data set. For example, the training data is 20% of data and validation data is 60% of data. Then, we averaged ten power rates to see which method generate the highest rate. We found out that Logistic regression is the best model. The averaged power is 0.562.

|  |  |  |  |
| --- | --- | --- | --- |
| Naïve power | Logistic Power | Tree power | K-NN power |
| 0.535529 | 0.5618905 | 0.5443745 | 0.516107 |