Determining the Vital Characteristics for Predicting the Number of Rings of Abalone Shell for Biological Research Using JMP® Pro 11

Ganesh Kumar Gangarajula, Yogananda Domlur Seetharama
MS in MIS with SAS & OSU Data Mining Certificate, Oklahoma State University, Stillwater

Abstract

Abalone is a species of sea snails. Their outer shell covering is a source of 'Mother of Pearl', which is used in jewellery, buttons, buckles and other decorative items. Abalone shells due to their amazing strength are being studied for developing next generation protective armour[1]. The strength of the abalone shells is directly related to the number of rings over the shell. The traditional process of understanding the strength of an abalone shell is by studying the number of rings on individual shell under a microscope after cleaning and processing. When dealing with large number of shells this process is ineffective as it is time consuming and tedious. Hence, predicting the number of rings on the shell based on its physical characteristics reduces the effort in determining the strength of the abalone shell.

Objective

The dataset gathered from UCI Machine Learning Repository[2] consists of 4177 observations, 8 independent variables such as Height, Shucked Weight, Shell Weight, Gender, etc. The target variable '*Number of Rings*' is a continuous variable. Our objective is to use JMP Pro 11 to predict the number of rings on the shell based on its physical characteristics.

Data Cleansing and Partitioning

- Two new dummy variables 'Gender' and 'Infant' are created to denote Gender and Infancy respectively. The dataset does not have any missing values.
- The data has been randomly divided in to 70% training and 30% validation data.
- A dummy variable 'Validation Indicator' is designated to indicate values of 1(validation) and 0(training). By specifying the validation indicator, JMP Pro 11 will distinguish between the validation and training observations.

Test for Collinearity

Multivariate collinear analysis using JMP Pro 11(Figure 1) revealed collinearity among:

- Shucked Weight (Meat Weight)
- Shell Weight (Weight of the dried shell)
- Whole Weight (Weight of the whole abalone)
- Viscera Weight (Gut Weight after bleeding)

Variable Selection

Shucked Weight, which is the actual meat weight of the abalone is a more accurate indicator of the abalone shell weight as it is not influenced by factors such as external build up, amoeba accumulation, etc. Hence, Shucked Weight is selected as an independent variable out of the four weights in order to predict the number of rings of abalone shells.

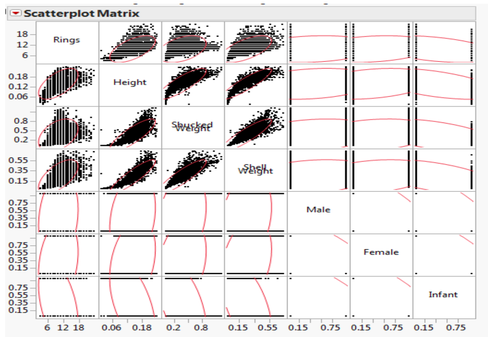

Figure 1: Multivariate Graph

Model Building

Regression Model

Using the power of JMP Pro 11, a Mixed Stepwise Regression model is built. As a stopping rule, p-value threshold is used with a cut-off value of 0.5 to enter or leave the model(Figure 2). Full factorial variable interactions to degree of 3 are applied to consider relationship between multiple input variables towards the target.

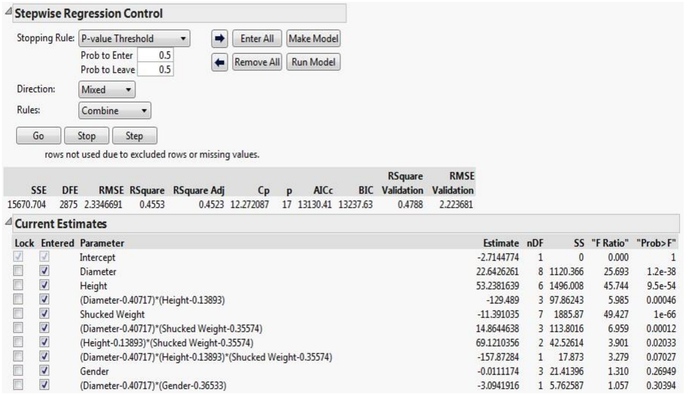

Figure 2: Mixed Stepwise Regression

Determining the Vital Characteristics for Predicting the Number of Rings of Abalone Shell for Biological Research Using JMP® Pro 11

Ganesh Kumar Gangarajula, Yogananda Domlur Seetharama
MS in MIS with SAS & OSU Data Mining Certificate, Oklahoma State University, Stillwater

Neural Network Model

The Neural Network model is built with 3 hidden layer nodes with Gaussian Activation Function. To offset the effects of input, response outliers and skewed distributions, the fitting options Transform Covariates and Robust Fit are specified in the model. A Squared method penalty function is applied to offset the over fitting of data by the model.

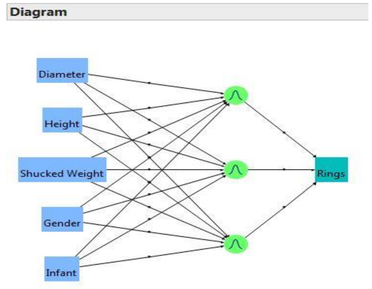

Figure 3.1: Neural Network

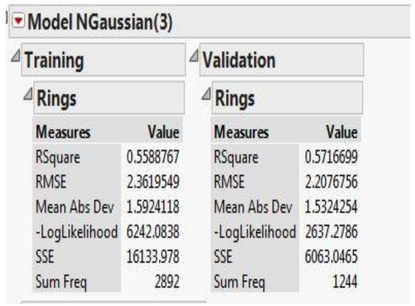

Figure 3.2: Model Performance

Actual Vs. Predicted Plots

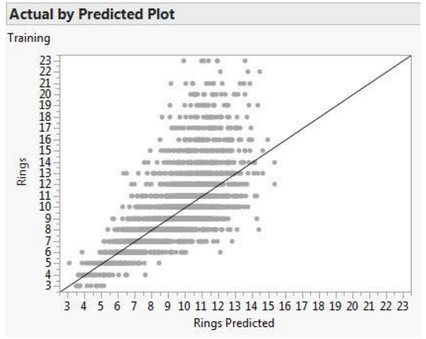

Figure 4.1: Training

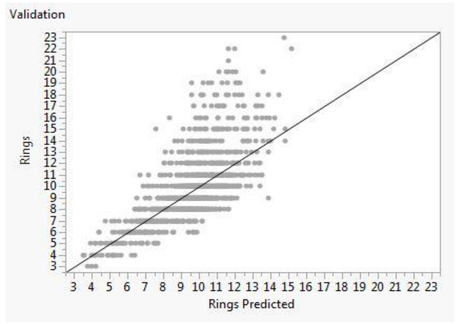

Figure 4.2: Validation

Results

According to the Stepwise Regression Model, the significant predictors for the number of rings are Diameter, Height, Shucked Weight and interactions between Diameter, Height and Shucked Weight of the shell. The Mixed Stepwise Regression Model has R square value of 0.45 on training data and R Square value of 0.47 on validation data(Figure 2).

As per the Neural Network model, the most important variables in determining the number of rings in decreasing order of importance are the Diameter, Shucked Weight and Height of the shell. The Neural Network model has R square value of 0.55 on training data and R Square value of 0.57 on validation data(Figure 3.2).

Conclusion

The Neural Network model performs better on the training and validation datasets as compared to the performance of Mixed Stepwise Regression Model. Hence the best model for predicting the number of rings on abalone shells is the Neural Network Model.

Acknowledgements

We thank Dr Goutam Chakraborty, Professor of Marketing and Founder of SAS & OSU Data Mining Certificate Program-Oklahoma State University for all his support and guidance throughout the project.

Reference

1. http://www.aps.org/publications/capitolhillquarterly/201105/seasnails.cfm
2. http://archive.ics.uci.edu/ml/datasets/Abalone [Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA]
3. http://support.sas.com/documentation/onlinedoc/jmp/11/UsingJMP.pdf

Disclaimer

We have analysed this topic using standard data mining and statistical techniques and we do not claim to have any kind of expertise in understanding the biology of Abalone Shells.