ABALONE DATASET ANALYSIS

Link to the data set source: https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/

Introduction:

Abalones comes from the family "Haliotidae", and these are marine snails with several respiratory pores on its convex like structured shell. The most number of abalones are found in the cold waters along the coastlines of, for example, Australia and Japan. Abalones are considered as a form of diet in various cultures all over the world either cooked or raw. It's relatively a lengthy process for abalone to enter into the market place. The meat from the Abalones can enter into markets freshly or canned or dried. Canned is the most preferred way while the dried will be least preferred form. On the other hand, shells after harvesting are used for décor, design of fret boards, jewellery because of its shining appearance as well as a currency for tribes.[1] [2] [3]

There are a number of recognised abalone species worldwide ranging from 30-130. The size of the abalone varies between 20mm to 200mm while the colour varies according to variety of species. The cost of Abalone has been increasing due to its popularity. On note of that, it is considered essential for the people to know the age of the abalone for getting on the price. The prediction of the age of these Marine snails is a hectic process. It includes cutting the shell and staining it. After that, the number of rings are counted manually under the microscope by the laboratory technician. Predicting the age of the abalone by the above mentioned method might be a tougher and time consuming task for either farmers or sellers. An innovative way to predict the age of the abalones has been identified by the researchers which is to utilize and understand the physical characteristics of the abalones for getting the age of the abalones. [1] [2] [3]

Description of the Dataset:

4177 observations of the Abalones have been collected. Each of the observation includes its 8 physical measurements and an anticipated number of rings. It is mentioned that adding 1.5 to the number of rings which is a count data will give the age of the Abalone in years. On moving to the physical characteristics of the Marine Snails, the shell length, diameter of the shell, and height with meat in the shell are in milli meters and are continuous variables. In addition, the weight of the complete abalone, the weight of the meat only (called Shucked weight), the gut weight after bleeding (called Viscera weight), the weight of the shell that is after dried are in grams and are continuous variables as well. A categorical variable giving the sex of the abalone is given. On investigation, it is seen that there are no missing values from the dataset chosen from the website. It is also said from the source that the attributes with continuous values are scaled by dividing by 200. [4]

Research Question:

Predict a model to determine the number of rings in the shell to acquire the age of the abalone as well as encounter the main relationships with the physical measurements incurred during this process.

Methods:

The abalones dataset has been checked for the missing values. Various methods have been used from analysing the dataset to the prediction of the age (research question). To start with, an exploratory analysis has been carried out for the variables to investigate on the relationships between the variables from the dataset using various types of plots and correlation plots. For correlation plot, data cleaning has been performed.
Delving into the analysis and approaching the research question, the complete dataset including the Sex Variable is taken for every method. For this dataset analysis, in order to predict the age, Regression is the method that will be used. Regression gives the predicted outcome when certain inputs are given. The basic linear regression equation will be

$$\mathbf{Y = a + bx,}$$

Where a is the intercept parameter, b is the slope and x is the value given for the input.

Firstly, simple linear regression with the target variable as “rings” has been used.
Before that, to fit any prediction model, the main task will be to know what variables can be thought as important to include in the model. This is done using the simple linear regression. There will be a lot of combinations of models based on the variables given and it is considered hard to check each of the possible combination. On note of that, different types of automated variable selection methods have been performed for this dataset to find an optimal set of variables efficiently by comparing AIC or Adjusted R^2 or RMSE or likelihood of different models. These include:

1. Step-wise forward & backward selection: A forward step-wise variable selection method starts with no variables in the model, iterates by adding one variable at a time until all the variables are put into some model and compares the models to find the best model while the backward step-wise variable selection method starts with a model that includes all the variables from the dataset and stops the iteration of getting rid of variables when a null model is observed. This method returns the best model by minimising the AIC. [9] [10]
2. Regsubsets: This method uses exhaustive search to find the best model out of all the possible combinations of the model by minimizing the BIC. [9] [10]

Once the variables that are the best fit for this model are identified, the next task would be to observe how the models fit or predict the data by checking the residuals. This was performed using two different methods

1. Cross validation approach: This method divides the dataset into 70:30 proportion of training and test sets randomly. The prediction using the simple linear regression has been carried out using all the variables from the dataset as well as the using the variables only from the backward step-wise variable selection method on the

training subset. The root mean squared Error has been calculated based on the models fitted on the test set.

2. Bootstrapped Cross validation: This method takes 25 (default) groups of division of the dataset rather than a single division as done in the first method, and performs the fitting of the model leaving out the first group for the testing by Mean Squared Error. This iterates for 25 times and the average of these test errors will be the estimated test error value. This has been performed using all the variables from the dataset and the variables selected from the backward step-wise selection method. RMSE has been calculated on the predicted values with these fitted models and the test subset to assess the performance of the models. [9] [10]

In general, we aim for the smallest errors for the fitted model. This is when Bias and Variance comes into existence. Bias gives the accuracy of the predicted values while the Variance gives how the fitted values are spread away from the regression line. Regularisation has been performed for the dataset which will be beneficial for the model performance. On note of that, Elastic Regression has been performed on the dataset. Elastic regression is a combination of Lasso Regression and Ridge Regression. A small modification to the simple linear regression leads to the elastic regression. This type of regression automatically selects the optimal value for alpha and lambda by penalising the model variables.

1. Ridge Regression: A model that penalizes the attributes that have large coefficients when compared to others to reduce the complexity of the model. This is much like equating the coefficients to 0. The lambda parameter finds its optimal value by either minimizing AIC or BIC. The attributes are centred and scaled.
2. Lasso Regression: Alpha will be the penalty parameter which penalises the attributes to decrease the complexity of the model by reducing the absolute sum values of the coefficients of the model. Alpha will take in the value of 1.

Alpha in Elastic regression is the combination parameter of both the ridge and the lasso while the lambda will act as the penalty for shrinking of the variables. The predictions using the elastic regression after using the bootstrapped cross validation on the training dataset, the performance has been calculated using RMSE. [5] [8]

Continuing the same Simple linear regression, another way of getting the predicted age and its performance has been performed. This includes checking of p_value and the significance level while selecting the parameters and its coefficients to be included in the model. It started with the complete model with “rings” as the target variable. At the end, the prediction performance has been calculated.

Secondly, a famous type of regression for binomial outcomes is the logistic regression. For this method, the rings column in the dataset has been divided into two categories. If the number of rings are less than or equal to 15, they are considered as young while the others are termed as old. This method basically deals with the log odd probabilities of the likelihoods of falling into a particular group. A positive value gives that its more likely to happen while a negative is contra. They are transformed back to the original terms by using exponents. Further, binned plots are plotted for analyzing the average residuals and the average fitted values of each bin/category. In addition, to assess the performance of the model, ROCR curve has been plotted for the predicted outcomes by making a graph with sensitivity (true positive rate) against specificity (true negative rate). Confusion matrix has been tabulated for this method as well as the accuracy is noted. This complete logistic regression analysis has been performed on the dataset with all the variables and using the variables given from the stepwise backward variable selection other than sex variable. Likelihood ratio test has been performed between these two models. [7]

Thirdly, clustering based on the variable Sex has been performed on the data. Gower distance has been used for the same. Optimal k value has been found using the elbow method. The performance of the clustering has been checked by using the silhouette method.

Lastly, Poisson regression has been performed on the dataset with the target variable being rings which is a count data. Poisson regression has been performed on the complete dataset with all the variables as well as using only the variables from the stepwise backward variable selection other than Sex variable. Further, robust standard errors and p-values have been calculated for the analysis of the model. After analyzing both the models by chi-sq test, and getting exponents of the old estimates, p-value has been dropped and the exponential coefficients have been put instead of Robust SE's. Finally, the data has been predicted based after extracting the respective coefficients of the model's variables that was considered better and they were added to the original dataset. [6]

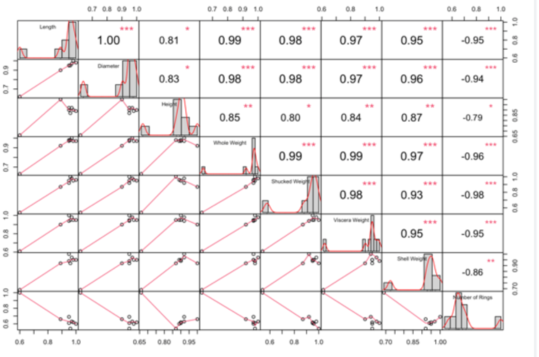
Data Cleaning:

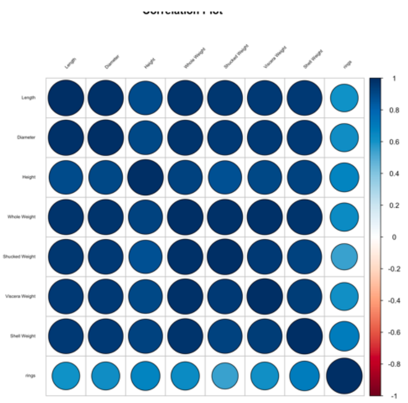
The column names of the dataset have been formatted to proper names for simple identification of the characteristics.
Further, to perform correlation analysis on the dataset, the categorical variable "Sex" has been removed temporarily. A subset of the data without Sex has been used to understand correlation between variables. For logistic regression, the rings column has been changed to a binary outcome, which will be 0 if the number of rings from the original data is less than or equal to 15 and 1 if the number of rings are greater than 15.

Exploratory Analysis:

Firstly, variable interpretation has been performed using different types of plots like Box Plots and Density plots. It is observed that the characteristic Length of the shell has a standard deviation of 0.12 with few outliers being less than the mean and the data is kind of normally distributed. The diameter of the shell also observes similar interpretation to that of the length of the shell. On the other hand, coming to other physical weights of the Abalones, few of the marine snails are weighed heavier than the mean weight. Comparatively, a higher deviation (0.5) from mean has been seen in regard with the whole weight of the abalone and Shucked Weight has standard deviation nearly less than half of that of Whole weight with others having nearly similar to that of the Length of the Shell. Also, the curve has a right tailed for each of the type of weights. The target data, Number of Rings, is somewhat normally distributed and few of the data predicted doesn't lie in the certain region where most of the data lies in with a standard deviation of 3.22. Coming to the Height of the shell, almost all the Abalones have similar height with few outliers. Other than the outliers for the height of the shell, the data is normally distributed. Further exploration tells us that all the characteristics except height of the shell observe an increasing trend with a slight curvature when compared to the Number of rings.

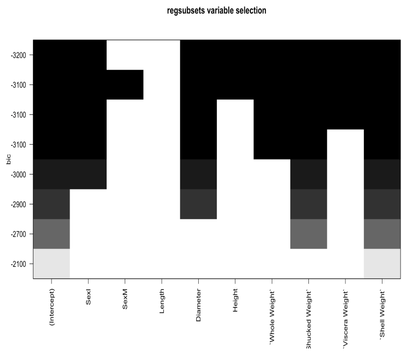
Secondly, there are three types of correlation that I considered to check for my analysis. Pearson Correlation, which works with normal and linear data, Spearman Correlation, which is a perfect fit with at least ordinal data and without any parameters, and finally Kendall Correlation, which checks on the strength of the dependencies between data. I have considered spearman correlation here as the target variable includes count data. [1] Well, there is not much of difference between Pearson Correlation and Spearman Correlation. In this case, from the interpretation of variables, we have seen that most of the data is not perfectly normally distributed and there is no linear relationship seen, Pearson Correlation has been put aside for this data and Spearman Correlation has been taken into consideration as it doesn't take the distribution of data into account. Also, as in count data, the order matters for the numbers for the number of rings, I would say that as spearman correlation has an underlying constraint that "if the data is at least ordinal" will be a better choice for analysing correlation between variables in this case. On analysing the correlation plot, different types of weight measurements came to have almost the same positive correlation with all other measurements except the height of the shell. The predicted variable, Number of rings have a positive correlation with each of the other variables, but they are not very highly correlated. From the correlation chart, taking an alpha value of 0.05, we can see that the relationship between the variables with three stars above the diagonal are considered significant.





Variable Selection:

Firstly, forward stepwise selection included all the variables in the model with the adjusted R^2 being 0.54 while the backward stepwise selection included Sex, Diameter, Height, Whole Weight, Shucked Weight, Viscera weight, and Shell weight with the same R^2 and residual standard deviation which is strange even though length has been excluded from the backward stepwise selection. Further, A regsubsets plot which is formed by penalizing the variables using BIC (lower BIC better) gives the same set of variables as given by backward stepwise selection which can be seen from the beside plot. The top one is considered as a better fit according to Regsubsets plot.



Simple Linear Regression:

1. **Cross validation:**
   As mentioned before regarding the importance of splitting of the dataset into training and validation set, the dataset has been split into 70% of training data where the model will be fit and 30% of validation set where the same model will be used to predict and check the performance of that model. On exploring the summary of the model fitted using the complete dataset and the variables from the stepwise backward selection for the training dataset, it is seen that there is not much difference in terms of either Adjusted R^2 or the deviance residuals. On calculation of RMSE, it is observed that in sample RMSE and out sample RMSE for both the models has been the same and the validation set's RMSE has been more than that of the training set indicating that the model has been overfitting.

2. **Bootstrapped Cross validation:**
   The model with the original dataset has RMSE as 2.21 and R squared as 0.533 while the simpler model has R squared of 0.527 and a higher RMSE which is kind of strange observation. Further, on calculation of the residual diagnostics (RMSE) of the predicted outcome and computing the residual diagnostics of the validation dataset, it is observed again that the original model has a lower Error than the selected model similar to the training dataset. The bootstrapped cross validation indicates that the model with the original dataset is a better model fit for the prediction of the age of the abalones.

3. **Elastic Regression:**
   All the variables from the original dataset have been centered and scaled. Using 25-fold bootstrapped approach and minimizing the RMSE, the alpha for the complete model came to be 1 and lambda as 0.004 which is near to be a lasso model. On extracting the coefficients using this model, it is seen that the coefficient of variable length has been shrinked to 0. As we know that analyzing the performance of the model after fitting one is as important as fitting the model, the RMSE for the validation data came out to be 2.286.
   On the other hand, when the number of folds is set to 10 and the elastic regression has been carried out, it is seen that the optimal value of alpha is 0.63 (much lesser than the before one) and lambda is 0.001 (similar to the previous one). Root mean squared error has been computed both for the training and validation set for the 10-fold repeated cross validation which came out that the model is over fitting- out sample RMSE higher than in sample RMSE.
   Surprisingly, the RMSE for the 10-fold and 25-fold cross validation varied by one-thousands only leaving the 25-fold repeated cross validation model a better one.
   As the abalone's dataset is not that highly dimensional, this might be the reason for elastic regression not showing better results as the elastic regression works better for models with greater number of predictors.

4. **Model using p-value:**
   The null hypothesis for the whole model that has been considered is all the predictors in that specific model will have coefficients 0 while the alternate

hypothesis will be that at least one of the predictors will have a linear relationship with the target variable. The second test will be the single test of a predictor where the null hypothesis will be equating the coefficient of that predictor to 0 and the alternate will be putting back into the model. The significance level is taken as alpha=0.05. For further comparison of models with different predictors, Adjusted R squared is considered. Initially, a complete model has been fit into the linear regression model. It is observed that we reject the null hypothesis as the p-value is less than the significance level giving us that there is a linear relationship between the target and at least one of the predictors. Further, checking for the single test for the variables Sex and Length, both have a greater p-value than alpha leading to fail to reject the null hypothesis and having those removed from the next model.
Next step included the first three significant variables Diameter, Height, and Whole Weight. It is observed that the adjusted R squared decreased. Using the same analysis for this specific model as well as for the single variable Whole Weight, it is seen that whole weight should be excluded from the next step of the model.
The third step would be to include the next variable which is Shucked Weight, and it is seen that it has a linear relationship with the target variable. Further, the same analysis has been done by including both the Viscera and Shell Weight in sequential order and came to a conclusion that Diameter, Height, Shucked Weight, Viscera Weight, and Shell Weight have a linear relationship with the target variable. But the final model using the single test says that the Viscera Weight has no linear relationship with the target variable. Even though the Viscera Weight has been excluded from the model, there is not much of a change seen in adjusted R squared between both the models. The performance has been calculated for the prediction and the error came out to be lesser than that of the previous models discussed.

Logistic Regression:

A logistic regression model has been fitted for the complete dataset as well as the selected model from the backward stepwise method other than the Sex variable. After exploring the outputs, it is observed that a comparatively lower AIC has been observed for the selected model and a greater range of deviance residuals. There is not much change in the residual analysis from the binned plots as similar number of categories have been observed outside the bands. There is not much of a difference seen between the two models with the ROCR curve as well. For both the models, the threshold value is taken as 0.1 with false positive rate being 0.1 and true positive rate being 0.7 for both the models. The accuracy of the prediction of models into old or young category is almost the same for both the models. A confusion matrix is also a similar one for both the models having 3710 observations classified correctly into the correct group for the complete model while 3709 observations correctly by the selected model. A likelihood ratio test has been performed for both the models and there is no significant model seen for this dataset.

Clustering:

On clustering based on Sex levels of the abalone’s dataset, it is seen that there is no one particular sex fit into one cluster on taking Gower distance because of the mixed type of variables in the dataset. The clusters are nearly of the same size but with a higher average

dissimilarity between the observations in the third cluster. Most of the Male abalones are put into the clusters 1 and 3 while a majority of infant abalones are in the second cluster. It is noticed that many of the female abalones are put into first and third clusters.
Further, both the elbow and silhouette method as well gives 3 clusters as a desired number.

Poisson Regression:

The complete model output starts with the function call and deviance residuals experience a very little bit of skewedness as the median is close to 0. Shucked Weight takes larger weightage in the regression with its coefficient equal to -1.8 which gives that the expected log count of rings is -1.8 for a unit increase in Shucked Weight. In the same way, diameter has the next positive weightage for the regression. Robust standard errors have been calculated for the model using the old coefficients and the covariance. This gave a better coefficient estimates for the model with the height error observed for the Height predictor. As we know that residual deviance measures the goodness of fit of the model, and we aim for a minimum residual deviance as a new model taking the variables from the backward stepwise selection method has been formed to test the models using chi-squared test. The degrees of freedom and the residual deviance has been increased in the second model when compared to the complete model. When the chi squared test has been performed to test the goodness measure fit, because of a lower residual deviance, the test has given the complete model to be a better model for this dataset. For further confirmation, the residual diagnostics have been performed on the complete data using both the models. For further confirmation, Root Mean Squared error has been calculated which gave a very negligible difference between both the models. Therefore, the complete model coefficient estimates have been exponentiated. Diameter predictor has the highest weightage for the prediction while the lowest will be Shucked Weight. The predictions of the number of rings have been finally put into the dataset.

Conclusion:

To conclude based on the different types of analysis performed on the abalone’s dataset, it is seen that the performance of the prediction of the number of rings of abalone is better for the original dataset no matter what model is considered for example simple linear regression or Poisson regression. Even though the logistic regression has been performed, the reason behind not finding any significant differences between different models for this dataset using the logistic regression can be understood from the type of the target variable, rings, which is a count data and the assumption of classifying the count data into two old and young is not a feasible idea. Further, from the clustering analysis, it is perceived that at least a few physical characteristics of male and female abalones are similar as the clusters 1 and 3 are divided equally within them. Finally, Poisson regression model with all the predictors has come to be the better model for this dataset to predict the age (in years) of the abalones which will be an addition of 1.5 to the predicted number of rings by the Poisson regression model.

References:

1. https://en.wikipedia.org/wiki/Abalone
2. https://luxuryviewer.com/why-is-abalone-so-expensive/
3. https://www.britannica.com/animal/abalone
4. https://archive.ics.uci.edu/ml/datasets/abalone
5. https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r
6. https://stats.oarc.ucla.edu/r/dae/poisson-regression/
7. https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html#assessing-logistic-model-fit
8. https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net
9. https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html
10. https://bookdown.org/jefftemplewebb/IS-6489/model-fitting.html#general-rules-for-variable-selection

Note: The graphs and outputs for each of the model are taken from the R code file.