Models for Predicting Abalone Sex and Value

Exploratory Analysis

Firstly, we will look at an overall summary of the variables in the abalone data. This will give us an idea of how the variables are distributed, and if there is a possibility that we have to some preprocessing before fitting our models.

Looking at the summary statistics we note that standard deviation between variables ranges quite significantly, for example standard deviation for abalone length is 24.02 while standard deviation for abalone height is 8.36. This indicated that we may have to transform some of the variables in order to ensure that there isn't a bias towards those variables during analysis.

Running a ggpairs plot we can see correlations, histograms and dot plots of each variable. A summary of findings is below:

- The dataset contains roughly similar quantities of each category. This is good as it means that there won't be much of a bias in classifying one category or the other.
- Strong positive correlations between variables
- Females and males overlap quite a bit in distribution while Infants has less overlap. Indication the differentiation between Male and Females might be tough.
- Some outliers found in for Male and Females as seen in the lower diagonal scatter plot.

Preprocessing

From the exploratory analysis we noted that variances range quite a bit between length and height and diameter and height. We will apply a square transformation on the length and diameter variables to help normalize the longer tail of these variables.

We also eliminate the outlier effects on the dataset to remove their influence on classification.

General Sex Classification

In the first part of our analysis, we are interested in classifying the sex of the abalone based on the dimensions of height, weight and diameter.

We begin by looking at the multi-class case, predicting the 3 classes altogether. The method we used to choose the appropriate model was to test 4 classifiers on the accuracy they produced. The classifiers we used was a linear discriminate model, a quadratic discriminant model, a radial support vector model and a linear support vector model.

The output of the accuracies for the 4 models is below:

Model	Accuracy
LDA	0.525
QDA	0.519
SVM radial	0.533
SVM linear	0.529

We can see in the above table that accuracies are generally low for the multi-class model. This is likely due to the high overlap in distribution that we found between males and females.

Classifying Infant Abalone

Next, we looked at the Bivariate case of Infants vs non-infants. The outputs that were produced for accuracy are much higher than they were for the multi-case. As we mentioned earlier, there was a much clearer distinction between infant distributions compared to males and females, and this has resulted in less confusion on our model.

Model	Accuracy
LDA	0.8
QDA	0.793
SVM radial	0.803
SVM linear	0.806

Classifying Female Abalone

We now look at classifying female abalones using the same testing procedure as we did before. This produces the following results for classification accuracy of females.

Model	Accuracy
LDA	0.692
QDA	0.681
SVM radial	0.693
SVM linear	0.692

Classifying Male Abalone

Continuing in the same fashion, we produce the results below:

Model	Accuracy
LDA	0.63
QDA	0.62
SVM radial	0.64
SVM linear	0.64

Conclusion

We conclude for the above tables that the SVM models had the best classification values. The below table shows the full summary of what was selected.

Classification	Model Chosen
Multi Class	SVM radial
Infants	SVM linear
Females	SVM radial
Males	SVM linear

Based on the classification results produced above we feel confident we can make an impact on reducing the number of infants that we harvest. The results show that our models will work well at selecting the adults from the infants but struggle at differentiating males and females. To solve this

issue, we may need to investigate collecting further variables that make up males, females and infants.

References

Medium. (n.d.). Medium. [online] Available at: https://medium.com/swlh/abalone-sex-classification-via-physical-measurements [Accessed 8 Dec. 2022].

Gandhi, R. (2018). Support Vector Machine — Introduction to Machine Learning Algorithms. [online] Towards Data Science. Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

www.datascienceblog.net. (2018). Linear, Quadratic, and Regularized Discriminant Analysis. [online] Available at: https://www.datascienceblog.net/post/machine-learning/linear-discriminant-analysis/#:~:text=Linear%20discriminant%20analysis%20(LDA)%20is.