Assignment 3: Executive Summary

Andrew L.J. Rezendes

School of Technology and Engineering, National University

DDS-8550: Predictive Modeling

Dr. Alexander Watt

June 29, 2025

Executive Summary

This project employed supervised machine learning models to predict the biological age of abalone, marine gastropod mollusks belonging to the family *Haliotidae*.Biological age of the species is approximated by the number of shell growth rings, using data from the well-known *Abalone* dataset (Nash et al., 1994). The predictive task focused on two models. Linear Regression was selected for its interpretability and usefulness as a baseline estimator, while Random Forest Regression was chosen for its ability to capture nonlinear relationships and mitigate multicollinearity (Breiman, 2001). Evaluation of the models was assessed by calculating Root Mean Squared Logarithmic Error (RMSLE) on each model to accommodate the right-skewed distribution of the target variable and penalize large relative errors.

Preprocessing the data involved removal of invalid height values (height ≤ 0), standardization of column naming to a snake\_case format, and domain-informed feature engineering. Newly engineered features included volume (a geometric proxy for shell size), density (mass-to-volume ratio), shell\_ratio (shell mass relative to total weight), and shell weight (difference of weights 1 and 2). These engineered variables were guided by ecological insights from mollusk morphology (Palmer, 1992).

Since there were no known labels, a standard 80/20 train-validation split was used (Kuhn & Johnson, 2013), balancing data efficiency with robust model evaluation. Linear Regression achieved an RMSLE of 0.1681, while Random Forest achieved 0.1663. The small margin of improvement (< 0.01) between the two RMSLE scores indicates that most of the predictive structures were linear and well captured by the engineered features. Were the structure less linear, the random forest model would likely capture this difference. This highlights the effectiveness of preprocessing in narrowing the performance gap between basic interpretable linear models and more advanced complex ensemble methods. The study affirms that principled preprocessing and feature construction can allow simpler models to perform competitively in real-world regression tasks.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., & Ford, W. B. (1994). *The population biology of abalone (Haliotis species) in Tasmania. I. Blacklip abalone (H. rubra) from the North Coast and Islands of Bass Strait*. Sea Fisheries Division, Technical Report No. 48. https://doi.org/10.13140/RG.2.2.19874.63688

Palmer, A. R. (1992). Calcification in marine mollusks: How costly is it? Proceedings of the National Academy of Sciences, 89(4), 1379–1382. <https://doi.org/10.1073/pnas.89.4.1379>

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3