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 was approximated by the number of shell growth rings, using data from the well-established Abalone dataset (Nash, Sellers, Talbot, Cawthorn, & Ford, 1994). The predictive task evaluated three models: a regularization-informed linear regression, principal components regression (PCR), and a baseline Random Forest regression. These models were selected to balance predictive performance with interpretability. Evaluation was based on the Root Mean Squared Logarithmic Error (RMSLE), which is robust to skewed target distributions and penalizes large relative errors.

Data preprocessing included removal of invalid observations (e.g., height ≤ 0), standardization of column names to snake\_case, and domain-informed feature engineering. Engineered variables included volume (a geometric proxy for shell size), density (mass-to-volume ratio), shell\_ratio (proportional shell mass), and shell\_weight (a derived weight differential). These features were informed by biological insight into mollusk morphology (Palmer, 1992).

Because no external labels were available, a standard 80/20 train-validation split was employed (Kuhn & Johnson, 2013), balancing data efficiency with generalizability. Among the models evaluated, the Lasso-informed linear regression achieved the best performance with a RMSLE of 0.1668, using a subset of nine predictors selected through L1 regularization. This model also demonstrated an R² of 0.602, suggesting substantial explanatory power. While residual diagnostics showed skewness and deviation from normality, the model's interpretability and sparsity make it a practical choice for ecological applications.

The Principal Components Regression model reduced the input space to three orthogonal components explaining 95% of the total variance. While this approach addressed multicollinearity, its predictive accuracy was lower, with a RMSLE of 0.1887. Additionally, interpretability was diminished due to the abstract nature of principal components and unclear variable loadings.

The Random Forest model, evaluated in preliminary experiments, achieved a RMSLE of 0.1663, slightly outperforming both linear models. However, the marginal improvement (< 0.01) suggests that the predictive structure of the data is predominantly linear and well captured by feature engineering. This aligns with previous findings that Random Forests perform robustly in complex domains (Breiman, 2001), but do not always offer substantial gains when simpler models are well-specified.

In conclusion, although Random Forests performed best in absolute terms, the Lasso-informed linear model demonstrated nearly equivalent accuracy with greater parsimony and transparency. This underscores the importance of combining domain knowledge with algorithmic strategy to design effective predictive models for real-world biological data.

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