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, a species of marine mollusks. 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 a regularization-informed linear regression and a principal components regression (PCR). These models are often selected to balance predictive performance with interpretability. Evaluation is based on the Root Mean Squared Logarithmic Error (RMSLE), which is robust to large magnitude differences in 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 difference of weights). These features were informed by biological insight into mollusk morphology (Palmer, 1992).

No external labels were available. As such, a standard 80/20 train-validation split was used, balancing data efficiency with generalizability (Kuhn & Johnson, 2013). Among the models evaluated, the Lasso-informed linear regression achieved the best performance with an RMSLE of 0.1668, using a subset of nine predictors selected through L1 regularization. This model demonstrated an R² of 0.602, suggesting moderate 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 components explaining 95% of the total variance. While this approach addressed multicollinearity, the predictive accuracy was lower than the linear model, with an RMSLE of 0.1887. Interpretability was diminished due to the abstract nature of PCR generating a linear combination of principal components. This makes it slightly more difficult to trace causality or biological meaning to the predictors.

Of the two models evaluated, the Lasso-informed linear regression achieved the best balance of accuracy and interpretability, selecting a concise set of biologically relevant predictors. While Principal Components Regression addressed multicollinearity, it sacrificed transparency by replacing original features with abstract components. These results highlight the importance of regularization and domain-informed feature engineering in building effective, interpretable models for biological prediction 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>

**Appendix A**

**Proof of Submission**

A screenshot of a video game

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.