Name: Vishu Wadhawan
Date: January 29, 2025
Course: STA 2453

Fish Classification Project Proposal

Introduction

Water ecosystems are considered vital to preserving global health but are notoriously difficult to explore. In an ideal economy, scientists would be able to monitor these ecosystems using underwater laboratories however, this has proven to be both costly and dangerous. Thus, recent research has developed new ways to learn about water ecosystems using sonar (sound). Sonar (**S**ound **Na**vigation **A**nd **R**anging) instruments emit sound waves into the water; if an object is in the path of these pulses, the sound bounces off the object and the reflected waves can be measured. The returning signal strength (in decibels) is traditionally measured by the following equation:

$$dB = 10 \times log_{10} \frac{P_{received}}{P_{reference}}$$

where $P_{received}$ is the power of the reflected sound and $P_{reference}$ is the omitted sound.

Project Description

This project will aim to classify three different fish species using sonar data. The three species of interest are 'lake Trout', 'lake Whitefish', and 'small mouth Bass'. The fish sonar data was collected by a group of researchers in Algonquin Provincial Park and their processed data set includes 14575 observations and 484 features. Each row (i.e observation) corresponds to a single fish from a particular species which equates to 7248 trout, 4285 whitefish and 3042 small mouth bass. Some notable features from the dataset include weight, girth, length, and the reflected signal strength (dB) from an emitted sound frequency. The emitted pulses in particular are labelled "F258" and "F259" ('F' for frequency) for example and they range from 46 dB to 259.6 dB; amounting to 425 different frequencies. One important aspect of this data set is that each fish is similar in size with respect to their specific species. As per figure 1, the length of trout ranges from 343 mm to 648 mm, the length of the whitefish ranges from 220 mm to 446 mm, and finally the bass lengths range from 190 mm to 479 mm. Moreover, there are a significant number of missing responses for certain sound frequencies, though no columns have entirely missing data. Specifically, the trout species has up to 2400

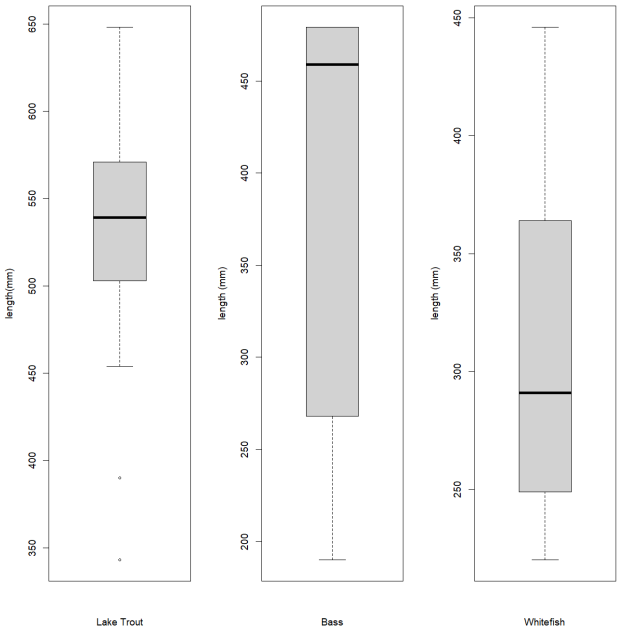
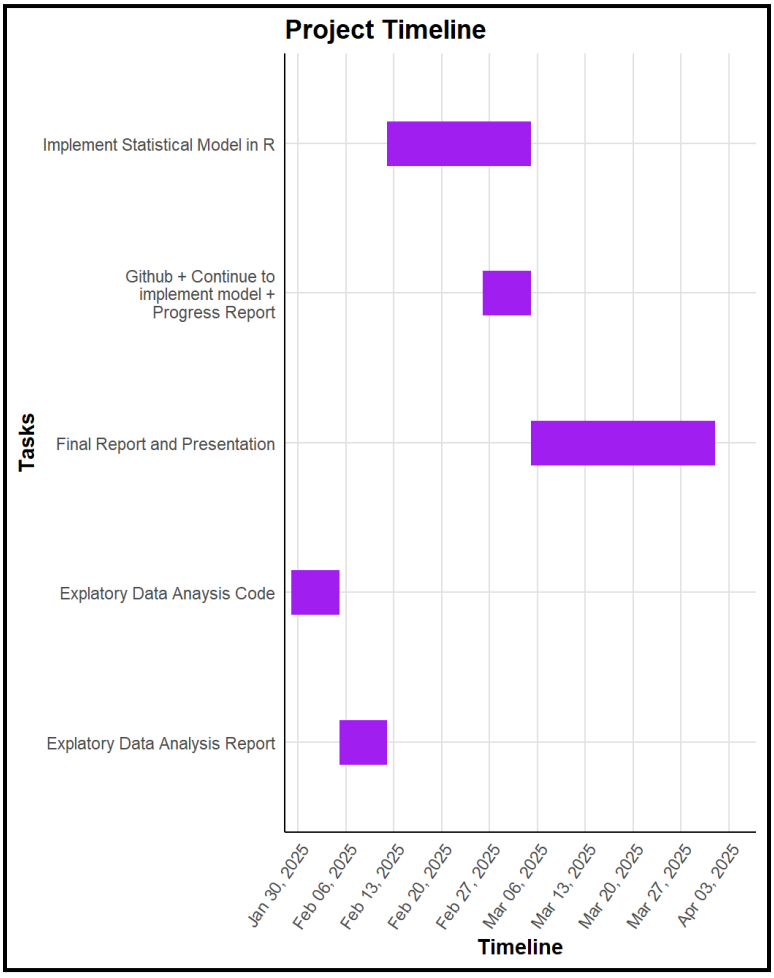


Figure 1: Showcasing the lengths of each fish species using boxplots in R.

missing entries for several sound frequencies while both whitefish and bass have <1000 missing entries in only a few categories.

For my classification model, I have decided to use only the reflected signal strength from each frequency as predictors and exclude the remaining variables in the dataset, as those features seem to be more closely linked to the physical features of the fish or the sonar instrument itself. As mentioned previously, all the fish approximately share the same phenotype with respect to their own species, so the remaining variables in the dataset are not expected to have an effect on the strength of the signal that is reflected back to the instrument. This decision ultimately shrinks the data set from the original 484 features to 425. Some potential classification models that can be considered are a decision tree, a basic neural network, or a multinomial logistic regression; albeit, each of these methods are limited by their own statistical assumptions. In theory, these models are well-suited for classification problems with very few classes and a large number of predictors which makes them feasible for this project. Finally, the accuracy of the model to be chosen will be calculated in a confusion matrix by means of k-fold cross-validation (CV). CV will help minimize the risk of overfitting by averaging results across multiple subsets of the data, ultimately leading to a more robust evaluation of the model's predictive performance.

Project Outline



Works Cited

James, G., Witten, D., Hasite, T., Tibshriani, R. (2021). *An Introduction to Statistical Learning with Application in R Second Edition*. Springer. https://doi.org/10.1007/978-1-0716-1418-1

Maher, R.C. (2014). *The Decibel Scale*. https://www.montana.edu/rmaher/eele417_fl14/decibel_scale_eele417.pdf

McCullough, H., Ambrose, B., Cromwell, M., Paulubicki, K., Sallis, A., Varner, J. (2020, March 30). *Understanding Our Ocean with Water-Column Sonar Data*. National Center for Environmental Information. https://storymaps.arcgis.com/stories/e245977def474bdba60952f30576908f?fbclid=IwAR1mEflE1Nk5285GRiqMtYC70Q-4z4xNT8H6HuKhUnsbMgqelXPpht8Nu2I