STA2453 Exploratory Data Analysis

Scarlett(Yi) Yang

February 2025

1 Introduction

The goal of this project is to classify fish species, Lake Trout (LT) and Smallmouth Bass (SMB), based on the features of sound waves that bounce back from the fish at different frequencies. This dataset contains 6,085 observations with 484 features and includes 3,828 observations for Lake Trout and 2,257 for Smallmouth Bass. The features include biological measurements such as total length and weight, and frequency responses from 45 Hz to 260 Hz in steps of 0.5 Hz. This EDA is used to determine how the frequency responses are different between the species and investigate which features are important for classification.

2 Initial Exploration

The original dataset from GitHub had more species, and some of the frequencies of the fishes were missing. After cleaning it up and filtering out other species, this cleaned dataset contains 9 Lake Trout and 7 Smallmouth Bass datasets with no missing values in frequencies. It is a little imbalanced: 3,828 observations of Lake Trout versus 2,257 for Smallmouth Bass.

3 Data Visualizations

3.1 Distribution of Key Variables

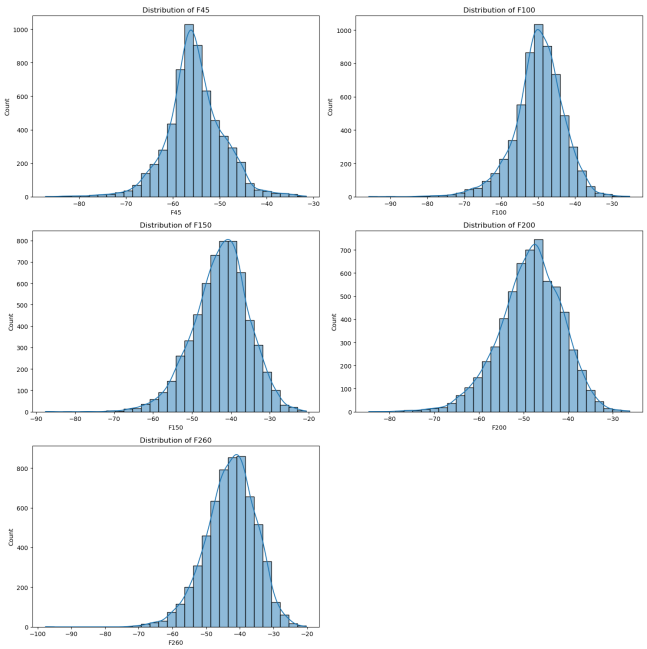

Figure 1: Distribution of Frequencies

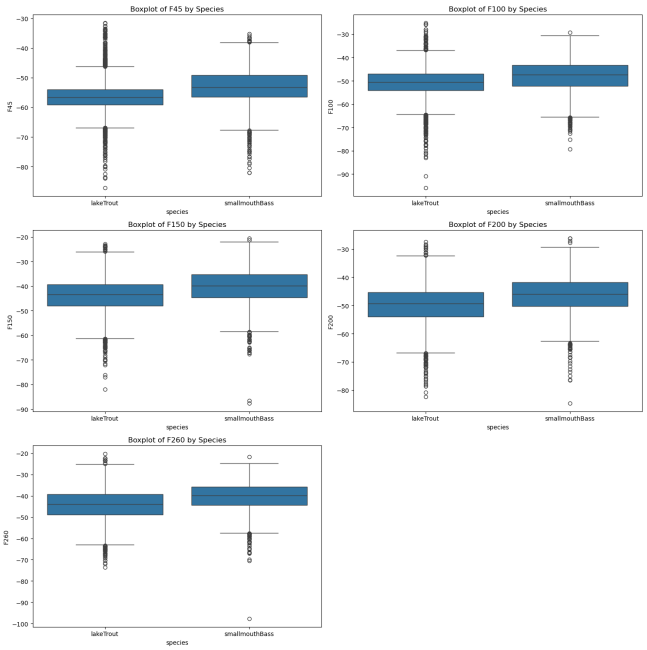

Figure 2: Boxplot of Frequencies by Species

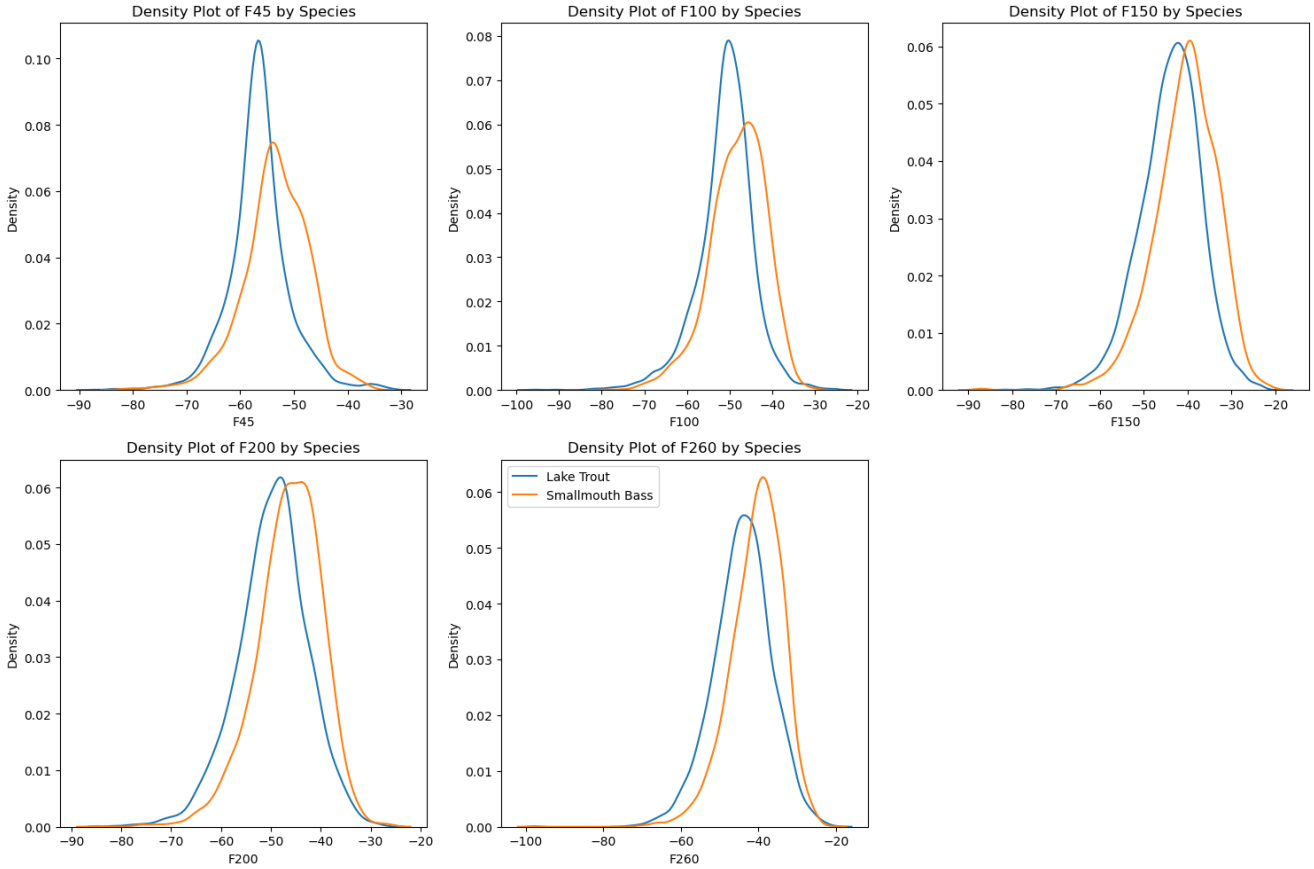

Figure 3: Density Plot of Frequencies by Species

F45 means frequency responses at 45Hz. Histograms of F45, F100, F150, F200, and F260 all indicate most values are within a certain range, but with some deviation. The deviations are consistent across frequencies and the center for high frequencies are higher in general. When comparing by species, boxplots and density plots both suggest that Lake Trout has much lower frequency responses compared to Smallmouth Bass. For Lake Trout, responses have a concentration on the low frequencies; for Smallmouth Bass, the strongest responses are at high frequencies. Most of the peaks are clearer in Lake Trout in density plots, exclusive of F260.

3.1.1 Time Series Analysis

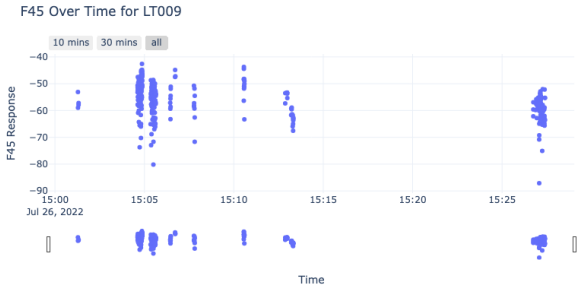

Figure 4: Time Series of F45 for LT009

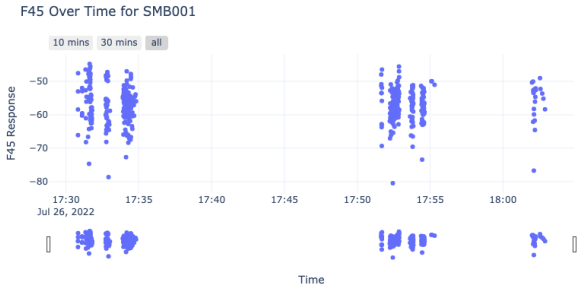

Figure 5: Time Series of F45 for SMB001

Time series-plot shows many gaps in data in the F45, we assume this also happens for other frequencies. In F45 time series-plot, Lake Trout has a greater frequency response variability ranging from -40 to -90, while Smallmouth Bass only ranged from -40 to -80. Interactive plots allow zooming in to see detailed trends and changes over time.

Missing records in the time series may have an impact on feature extraction and modeling. Considering this issue, species-specific imputation, resampling methods, feature engineering, and state-of-the-art modeling techniques which are capable of handling missing data must be implemented. This can give better accuracy of classification and also more insights about the behavior of Lake Trout and Smallmouth Bass.

4 Correlation Analysis

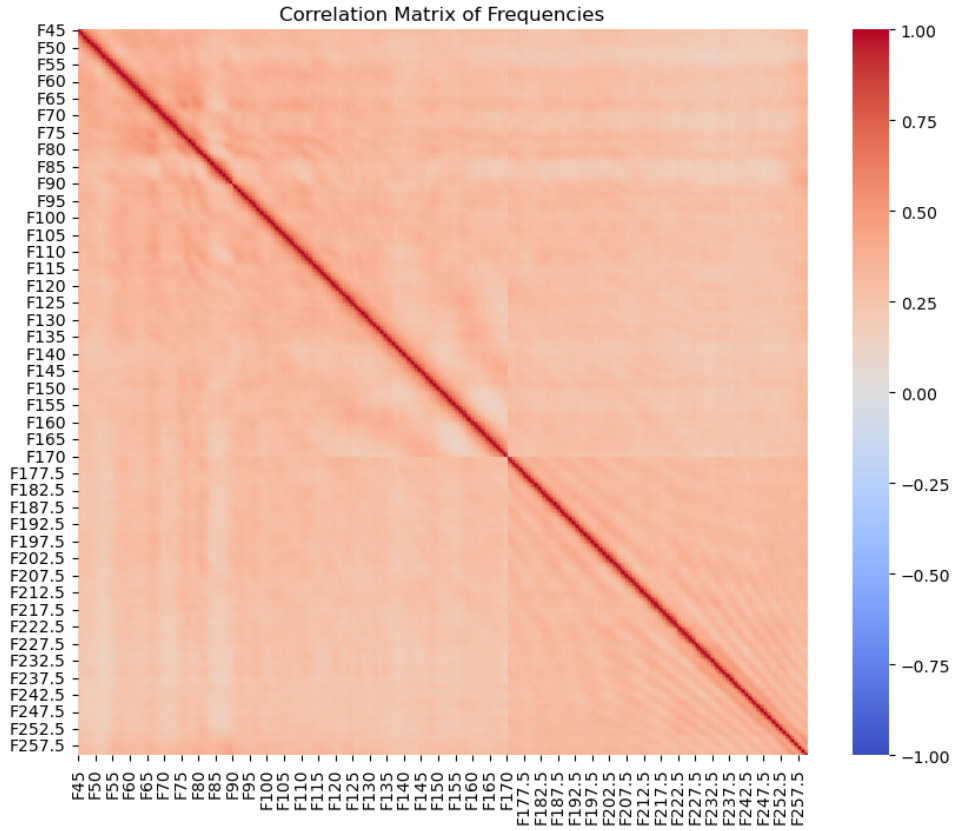

Figure 6: Correlation Matrix

The correlation matrix shows strong positive correlations between neighboring frequencies. For example, F45 and F45.5 are highly correlated. There are clusters of highly correlated frequencies, which may suggest redundant neighboring features. The heat map of the correlation matrix below shows that most frequencies are moderately correlated.

5 Species Comparison

5.1 Mean Comparison

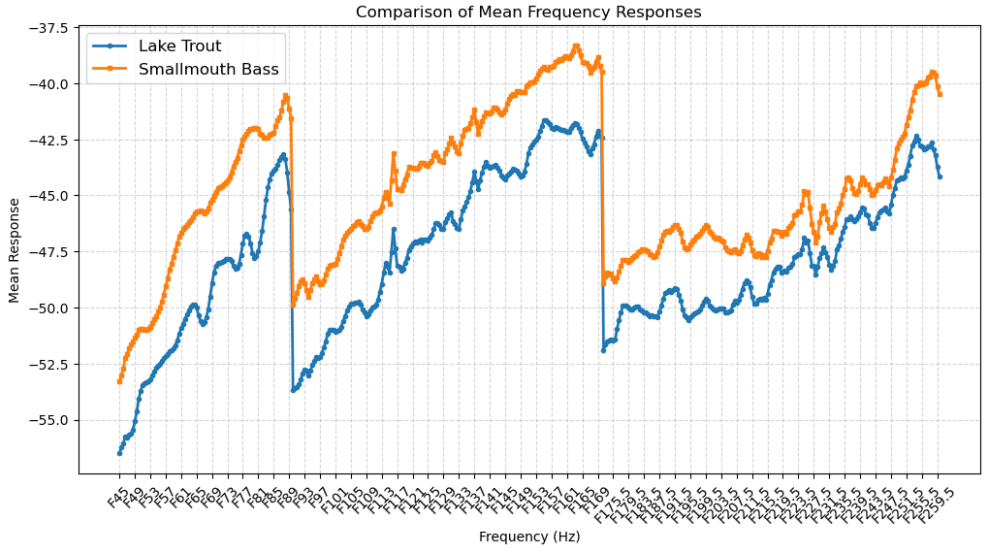

Figure 7: Mean of Frequency Responses

Frequency response means for Lake Trout and Smallmouth Bass are well separated, with Lake Trout lower than Smallmouth Bass in most of the frequencies, while the trend of frequency response is similar in both species. However, for classification, the absolute values of frequency responses are more useful than trends.

5.2 Feature Importance

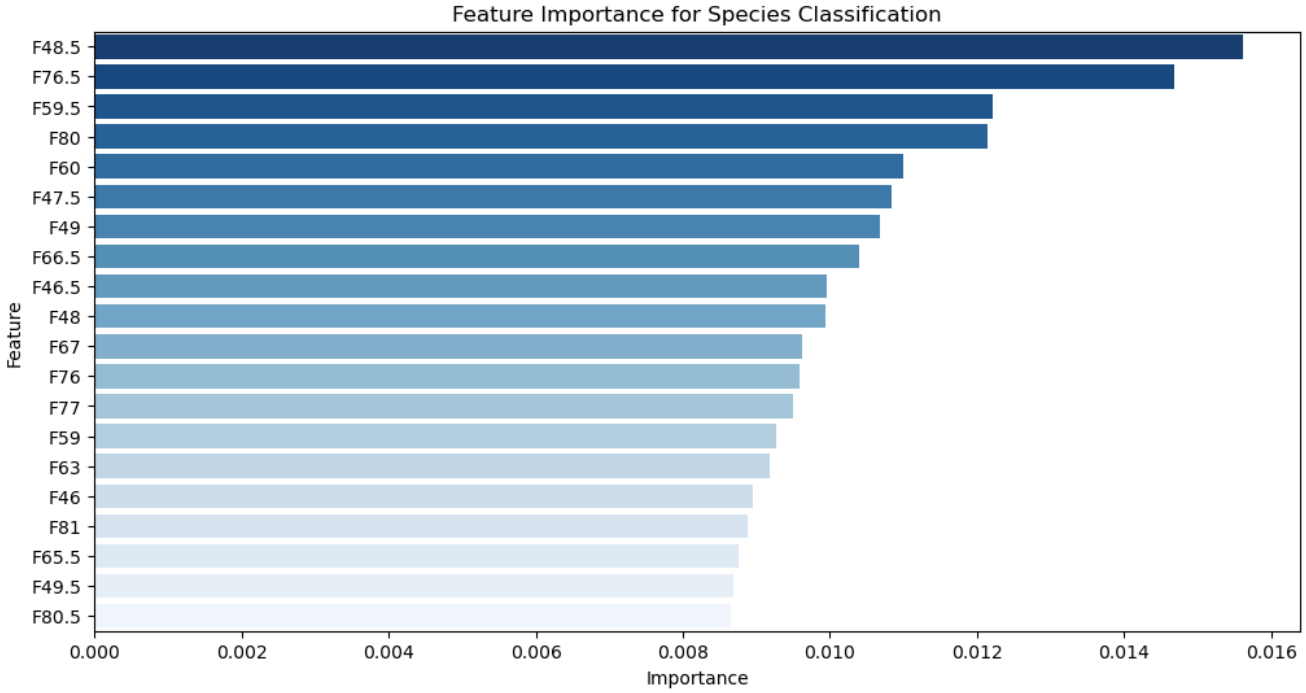

Figure 8: Feature Importance

F48.5, F76.5, and F59.5 are the most informative features for classification, as identified by the Random Forest classifier, though not significantly more so than others.

6 Conclusion

This exploratory analysis has pointed out the main differences in frequency responses between Lake Trout and Smallmouth Bass, the gaps in the dataset that may affect modeling, and very highly correlated features that could be reduced by using dimensionality reduction techniques. Further work will involve the application of PCA in removing redundant frequencies, addressing missing data in the time series, testing other classification models, and adding biological measurements to increase the accuracy.