Brandon Bernstein, Ho Lam Wan, Rohan Basavaraju, Ying Chen

AMS 380 Project (Group 8)

## Identify & Define the Problem Statement

This dataset is a record of 7 common different fish species in fish market sales. In this project, we will be using various regression models to predict the weight of the fish.

**Target Variable**: Weight: Weight of the Fish in grams (g)

**Predictor Variables**:

Length1: Vertical length of the Fish in centimeters (cm)

Length2: Diagonal length of the Fish in centimeters (cm)

Length3: Cross length of the Fish in centimeters (cm)

Height: Height of the Fish in centimeters (cm)

Width: Diagonal width of the Fish in centimeters (cm)

For innovation we use ridge regression.

## Collect, Clean, & Explore the Dataset

There were no nulls or notable incorrect measurements in the data set. With 159 complete observations we grouped the data set by fish species. Perch was the most popular species with a count of 56, followed by Bream which had a count of 35. All other species had 20 or less observations. The pike species in particular showed a different trend than the other

Figure 1

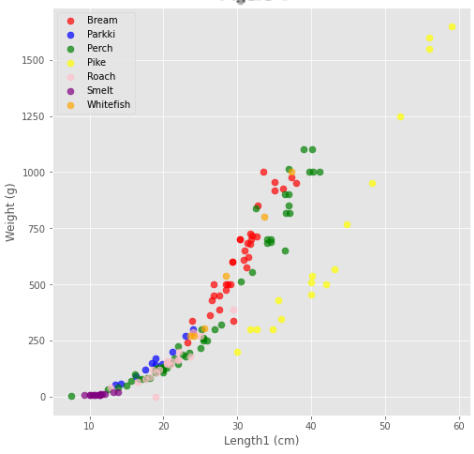



Figure 2

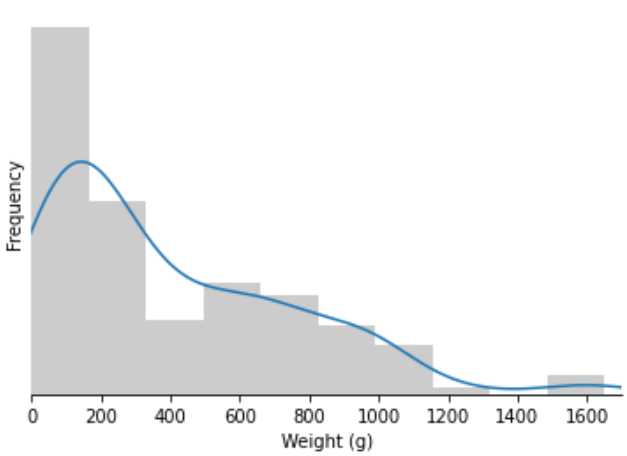



Table 1

| | Weight |
|---|---|
| count | 159.000000 |
| mean | 398.326415 |
| std | 357.978317 |
| min | 0.000000 |
| 25% | 120.000000 |
| 50% | 273.000000 |
| 75% | 650.000000 |
| max | 1650.000000 |

species in multiple categories, with 3 large outliers. Figure 1 shows this trend for Length1. Weight (g) also had a strong right skew in its distribution (Figure 2), with the summary statistics above in Table 1. However, we did not partition the data into sections with and without the pike species because we wanted to create an overall linear model. Instead a dummy variable was created based on each species.

**Analyze the Dataset & Measure Success**

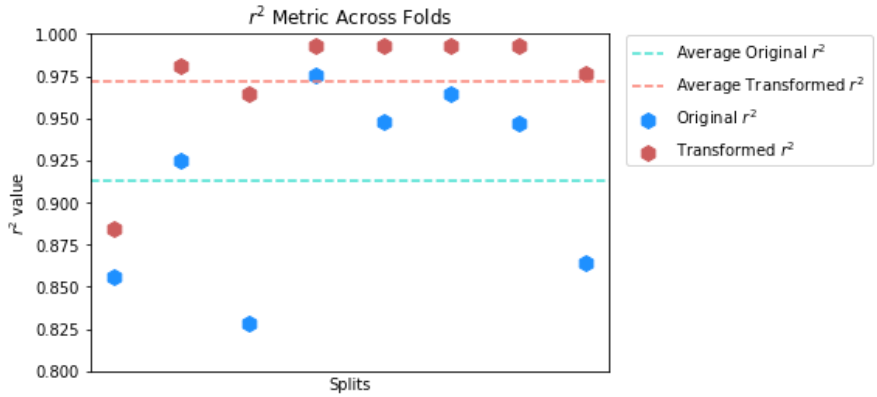
Multiple regression techniques were used in this project.

**Linear regression and Transformation**

As above-mentioned, we added extra dummy variables (1s and 0s) to indicate the species of data and aim at analyzing a linear relationship for all species. We split the dataset into 70% training data and 30% testing data. After carrying out the original model, based on the relationship between “Length1” and “Weight”, a square root transformation was applied on the dependent variable. Below is the table showing the result of models without (indicated as yellow) and with the square root transformation (indicated as blue).

| | R-square | MSE |
|---|---|---|
| Training data - all variables | 0.936 | 8139.1 |
| Training data - removed collinearity | 0.933 | 8213.6 |
| Test data | 0.884 | 11176.0 |
| Training data | 0.988 | 0.986 |
| Test data | 0.947 | 3.703 |

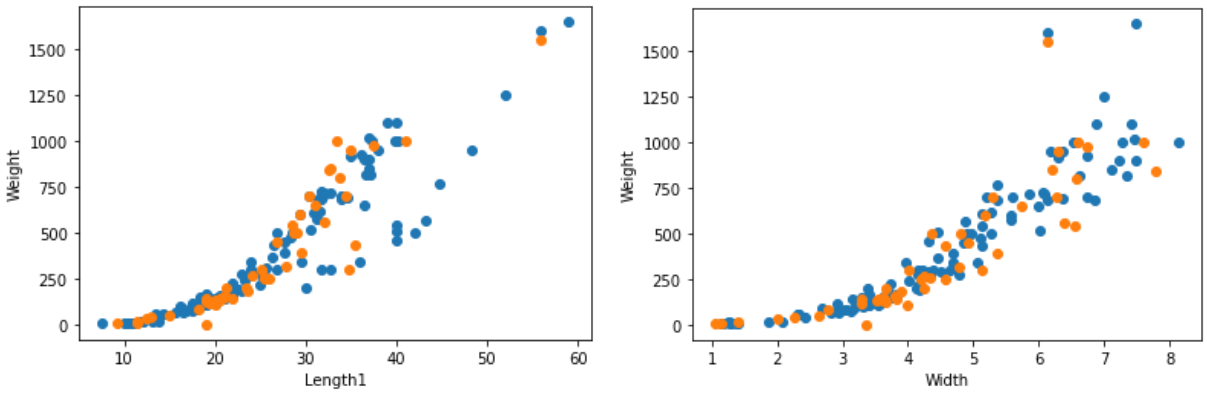
Cross validation with 8 splits are used:




The model improves after doing the transformation with over 90% in R-squared in both cases.

**Polynomial Regression**

We apply polynomial regression to capture the non-linear relationship between variables. Observe that both independent variables “Length1” and “Width” have good polynomial relationships with the target variable “Weight”. Two tables below show the relationship between independent variables and target variables in both the training set (blue) and test set (orange).



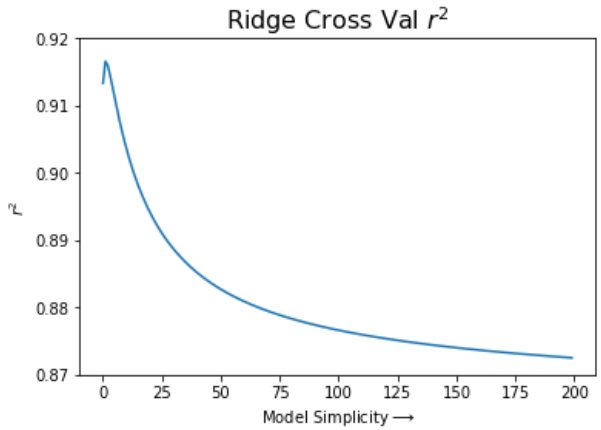

However the accuracy of the model is decreased due to the difference in species. Below is the summary table:

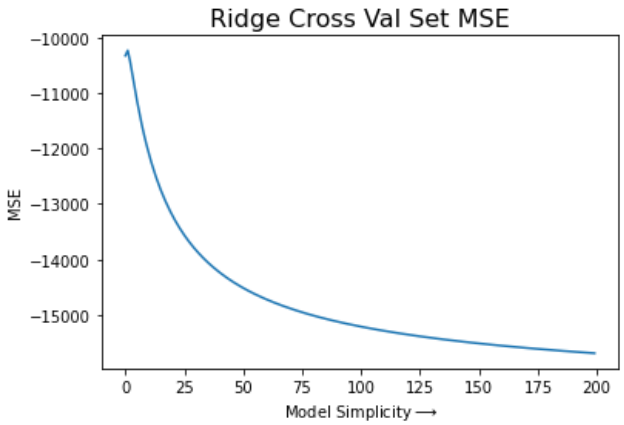
| | Degree | R-squared (Training set) | R-squared (Test set) |
|---|---|---|---|
| Length1 | 6 | 0.864 | 0.831 |
| Width | 5 | 0.853 | 0.774 |

The trade-off of polynomial models is that the outliers of the pike species greatly affect the performance of the models.

**Ridge Regression and Transformation**

Briefly speaking, ridge regression is a model we can use when there exists multicollinearity in the data. It adds a weight that simplifies the model dragging the slopes to 0. After implementing cross-validation with different choices of alpha for the loss function, it shows that the model performs better if alpha = 1 where both graphs achieve their optimal values.

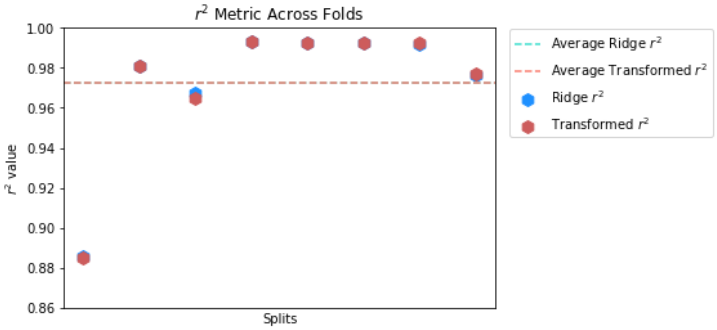


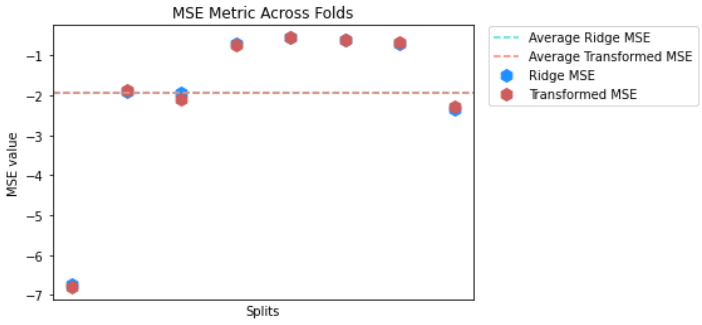




We apply the same transformation on the dataset as shown before - the R-squared increases!

| | R-squared (Training set) | R-squared (Test set) |
|---|---|---|
| Ridge without Transformation | 0.929 | 0.887 |
| Ridge with Transformation | 0.988 | 0.947 |

Compared to the original transformed (square root) model, we can see that ridge regression does not provide a significant increase in performance.



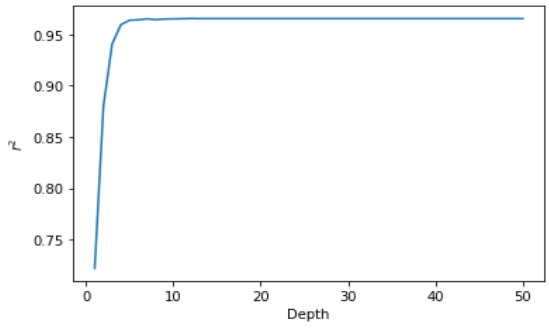
**Decision Tree and Random Forest Regression**

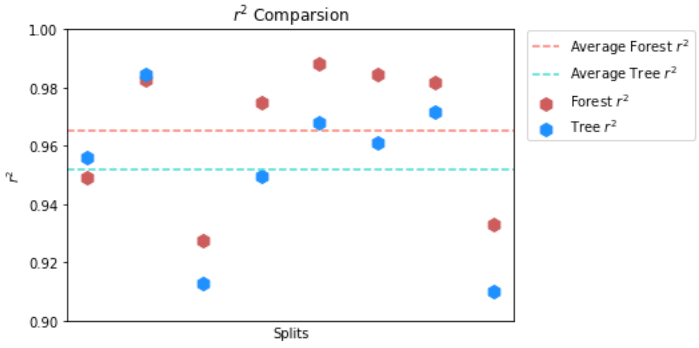
Decision Tree builds regression models in terms of a tree structure separating data into multiple leaves in each node. It uses the concept of entropy (information gain) when deciding the nodes. Random Forest is an ensemble machine learning method by constructing numerous decision trees to fit the relationship between variables.

In the Decision Tree method we will be using “30” as the max depth of trees; while in the Random Forest method we will be using “10” to give satisfactory R-squared results as the R-squared becomes stable around 10 (Figure 1).

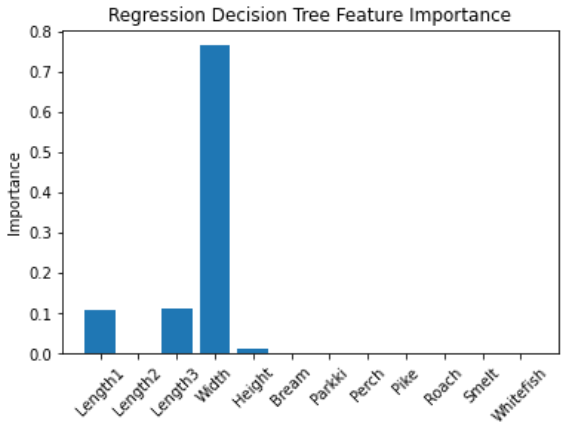
Cross validation scores (with cv = 8) are used in both cases to calculate the R-squared.

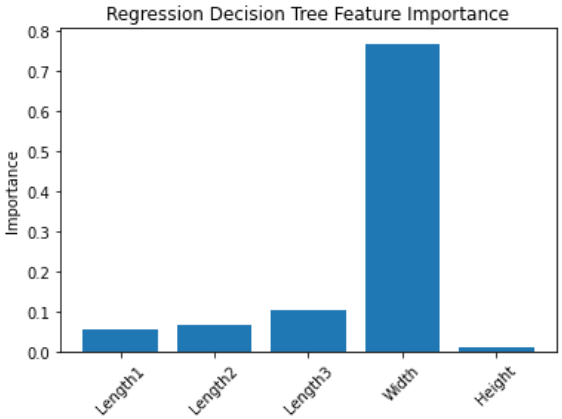
We can observe that Forest on average performs better than Decision Tree, with around 0.97 in R-squared value.




Both methods give “Weight” as the most influential variable, followed by “Length3”.



**Report Final Results**

Our main aim for this project is to utilize various models and see which variables best allow us to predict the weight of the fish. From our linear regression model, which tests the relationship between Length1, Height, Width, and the dependent variable Weight. We learn that the average $r^2$ value in our original model is relatively high (~ 0.91) but even higher after we perform a square-root transformation (~ 0.975). Our results suggest that linear regression is a good fit for this relationship.

From our polynomial regression, the results show that both “Length 1” and “Width” have a decent polynomial relationship with “Weight”. As seen from our graphs, when the two independent variables increase, Weight grows exponentially. While our $r^2$ values are slightly lower in this model as opposed to the linear regression model, it does not mean this model is not a good fit for our datasets. The accuracy might have been decreased due to outliers and differences in species of fish caused by the pike species.

From our ridge regression, as model simplicity increases, the $r^2$ and MSE values increases and then decays exponentially, peaking when alpha = 1. This indicates that with a more complex model, it would be a less accurate fit (lower $r^2$ and a large, negative MSE). After a square-root transformation is perform on this model, an increase is seen in our $r^2$ values compared to the original model.

From our decision tree and random forest regression, the average $r^2$ value is somewhat higher in the forest regression model in comparison to the decision tree, albeit not by much. Our results show that Width is the most important variable in both models when it comes to predicting the weight of the fish. Although in the graph with all the fish listed, it might seem like “Length 1” and “Length 3” have equal importance, we can very obviously see that “Length 3” is the 2$^{nd}$ most important variable after zooming in on the independent variables portion.