Jennifer Markham
Capstone 2: Crab Aging Project Report
Springboard: Data Science Bootcamp

I. Background/Motivation

Commercial crab farming is a popular business in coastal areas. The success of the business is driven by the popularity of crab consumption in many countries around the world. However, after a certain age, crab growth becomes limited and any size gains become negligible. In order to reduce cost and increase profits, it is critical for commercial crab farmers to know the optimal age to harvest crabs. A dataset was provided which contains data on ~50,000 crabs. The age and various physical attributes- such as Weight, Height, and Diameter- are recorded for each crab. The goal of this study is to predict crab based on physical attributes to help optimize crab farming practices.

II. Exploratory Data Analysis

Exploratory data analysis was performed prior to creation of a model to better understand trends in the raw data. Scatterplots of each Age versus each numeric physical attribute (Fig. 1) and a correlation heatmap (Fig. 2) were made to determine if any correlations existed between Age and a single physical numeric attribute. Pearson correlation coefficients between Age and each numeric were calculated to quantify any linear correlations, the strongest correlations were ~0.60-0.65 from the attributes Shell Weight, Height, and Diameter. The Pearson correlation coefficients between Age and each numeric attribute were also calculated on data subsetted by crab gender as well to quantify if any stronger linear correlations existed by Gender. Indeterminate gender data had the highest Pearson correlation coefficients between Age and numeric attributes, ranging from ~0.6-0.7 with the attributes Shucked Weight, Viscera Weight, and Shell Weight.

In addition, the distribution of each crab attribute was reviewed prior to modeling to ensure that the observations were indicative of the actual crab population. Gender observations were observed using a bar chart to ensure an even distribution between the 3 crab genders (Fig. 3). Each numeric attribute's distribution was reviewed using a boxplot and histogram (Fig. 4). Numeric attribute outliers were removed using the interquartile range approach described below. Data points for each numeric physical attribute that laid above the upper bound and below the lower bound as defined by the interquartile approach were removed from the dataset prior to modeling.

- Upper bound for outliers: Q3 + 1.5 * IQR
- Lower bound for outliers: Q1 - 1.5 * IQR

Finally, a principal component analysis was completed on the scaled data with outliers removed to determine if dimensionality of the dataset could be successfully reduced. The PCA analysis showed that dimensionality of the dataset could be reduced such that 2 principal components consisting of linear combinations of the crab features could account for ~97% of the cumulative variance seen in the dataset (Fig. 5).

III. Modeling

For this study, linear regression, decision-tree regressor, and random-forest regressor models were trialed. Prior to modeling, a feature selection was performed to determine which number of features should be included in the model (Fig. 6). Linear regression modeling was subsequently performed, trialing inclusion of each number of possible features in the dataset. 5-fold cross validation was performed for each linear model created (Fig. 7). The best performance was achieved using a linear model including all features, the summary of its performance metrics is below.

- Coefficient of determination: 0.54
- Mean absolute error: 1.20
- Mean square error: 2.47

Next the decision-tree and random-forest regressor models were also tried using default parameters and hyperparameter tuning with RandomSearch. Optimum performance was found using the default decision-tree regressor model, with mean absolute error for crab age of 0.119. The default decision-tree regressor model listed the weight attributes- Shell Weight, Shucked Weight, Weight, and Viscera Weight as those with the highest feature importances (Fig. 8).

IV. Conclusion

Accurate prediction of crab age will help crab farmers develop more profitable & sustainable crab farming practices. A decision-tree regressor model should be used to predict crab age based on crab attributes, which includes physical characteristics and gender. The default decision-tree regressor model developed can accurately predict crab age with a mean absolute error of 0.119. Crab weight attributes were found to be the most critical features in the decision-tree regressor model. Future work may include trailing more advanced tree algorithms such as XGBoost or completing alternate methods of hyperparameter tuning for the decision tree and random forest models to optimize performance.

V. Appendix with Figures

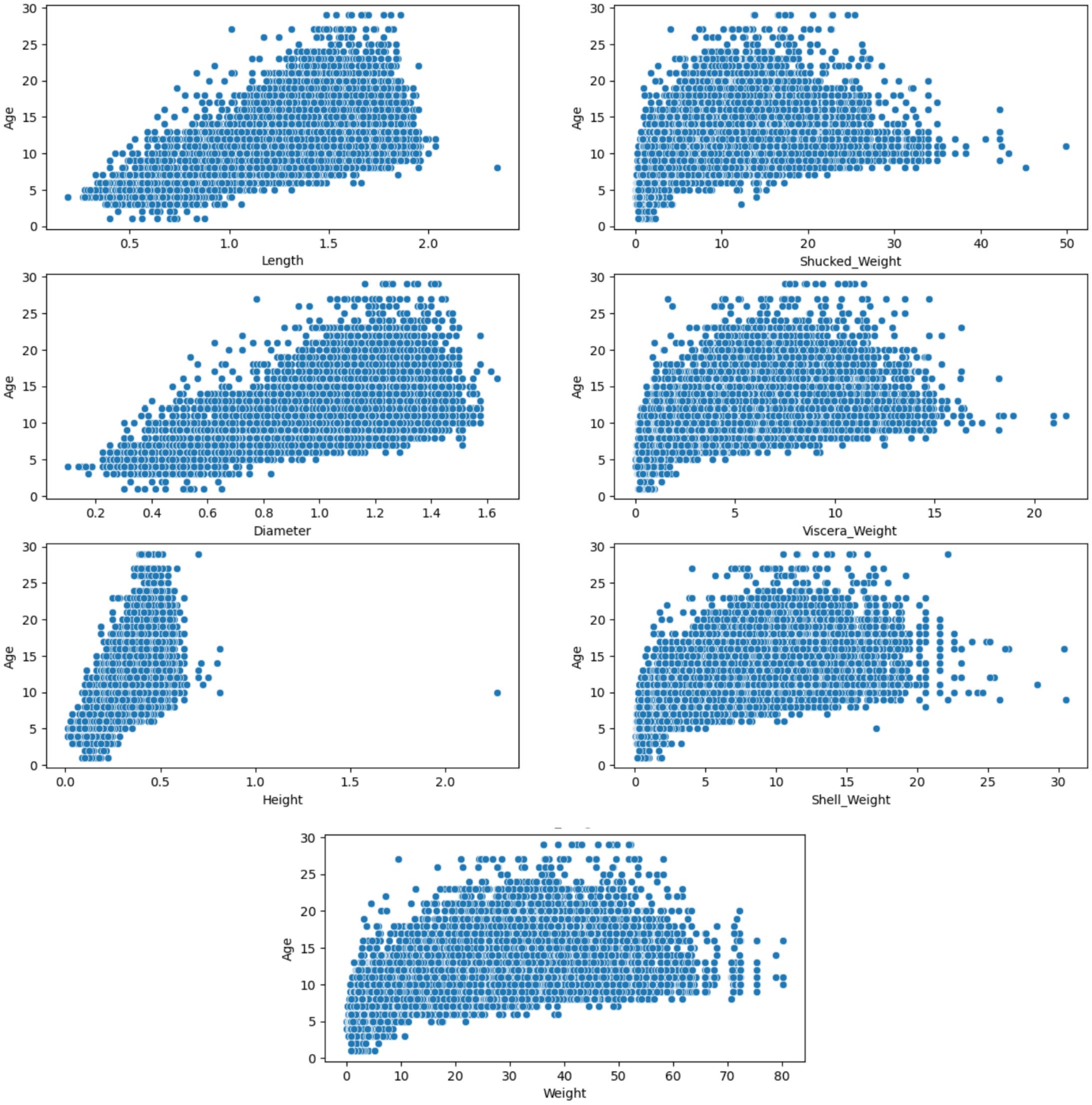

Figure 1. Scatterplots of Age vs each numeric attribute in the dataset.

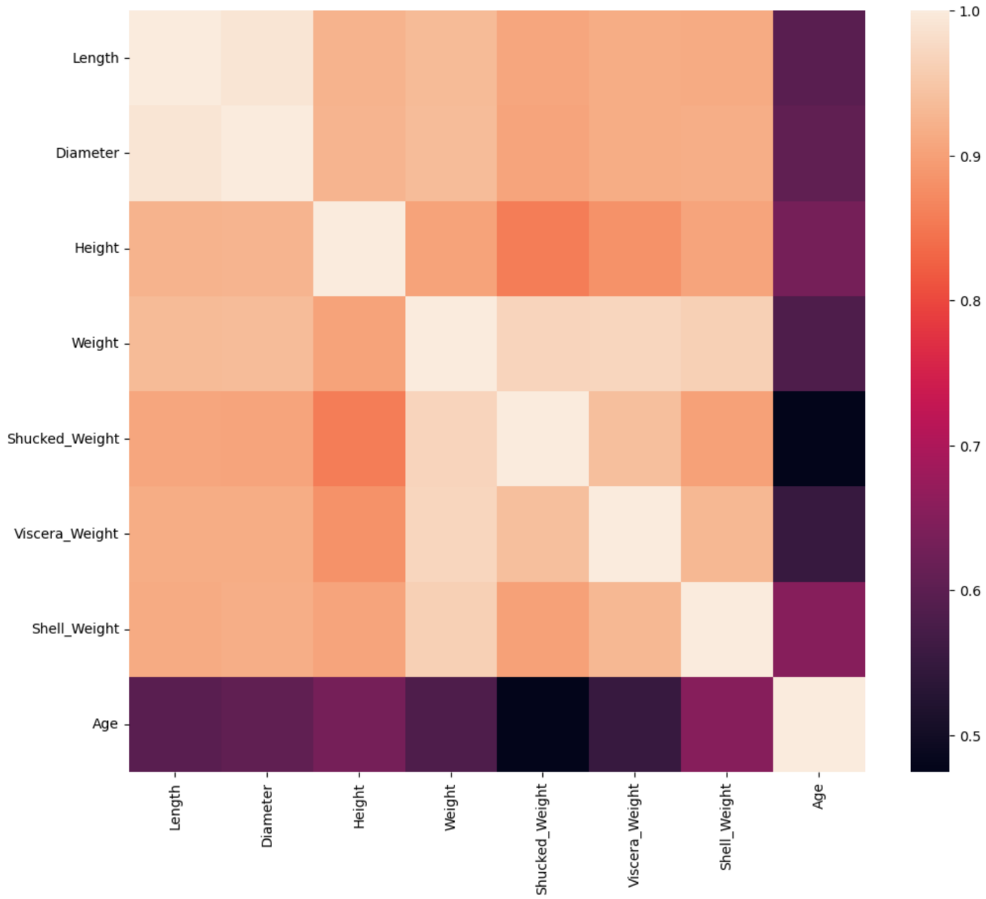

Figure 2. Correlation heat map of Age and each numeric attribute in the dataset.

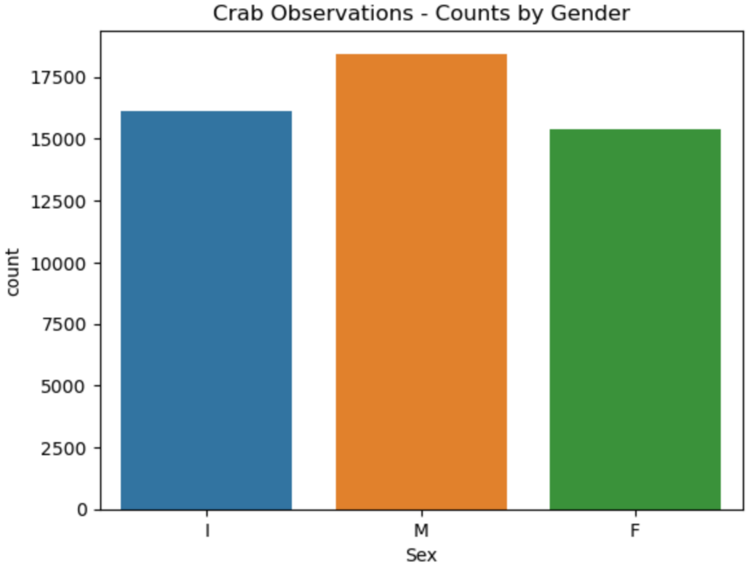

Figure 3. Distribution of gender observations in crab dataset.

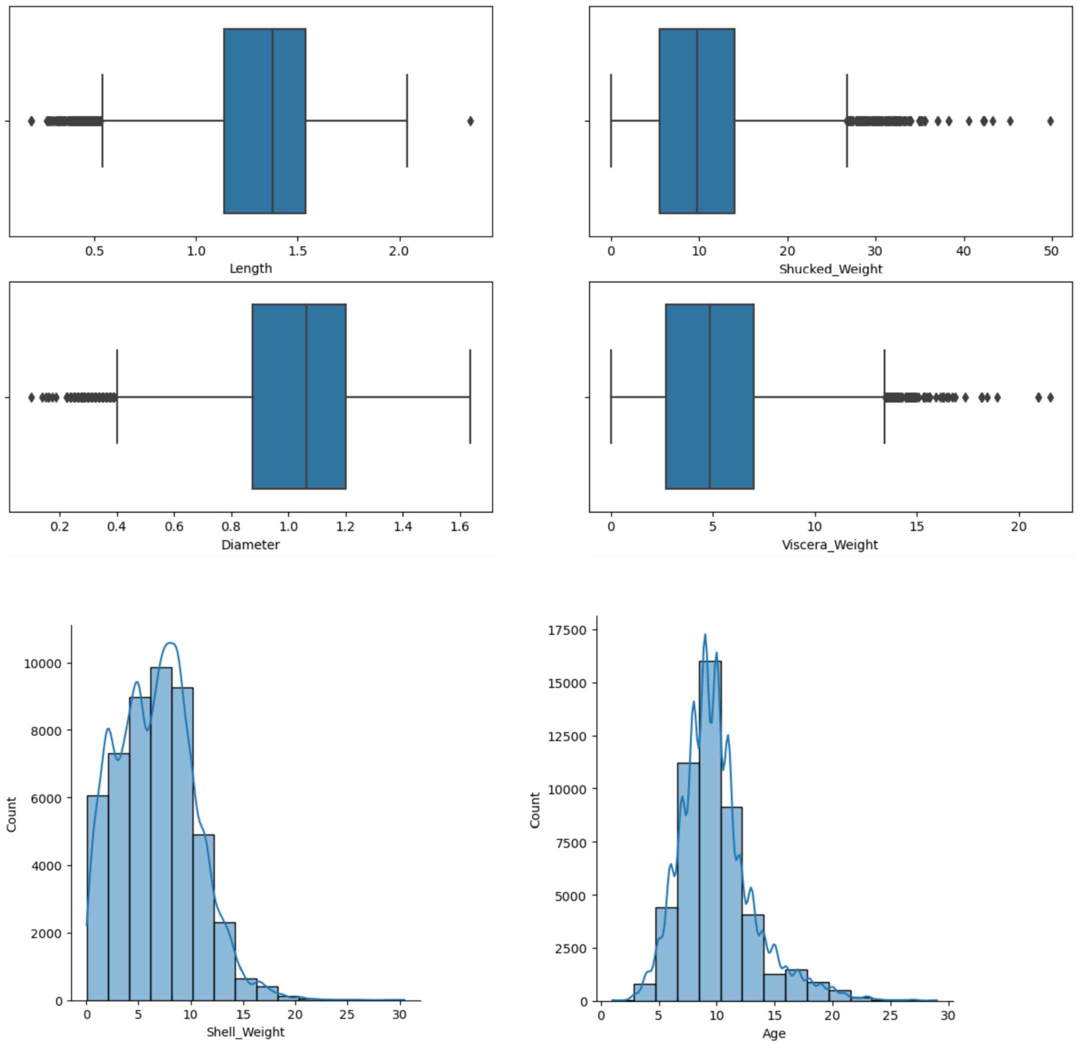

Figure 4. Each numeric crab physical attribute's distribution was reviewed using a boxplot and histogram. Examples are shown above.

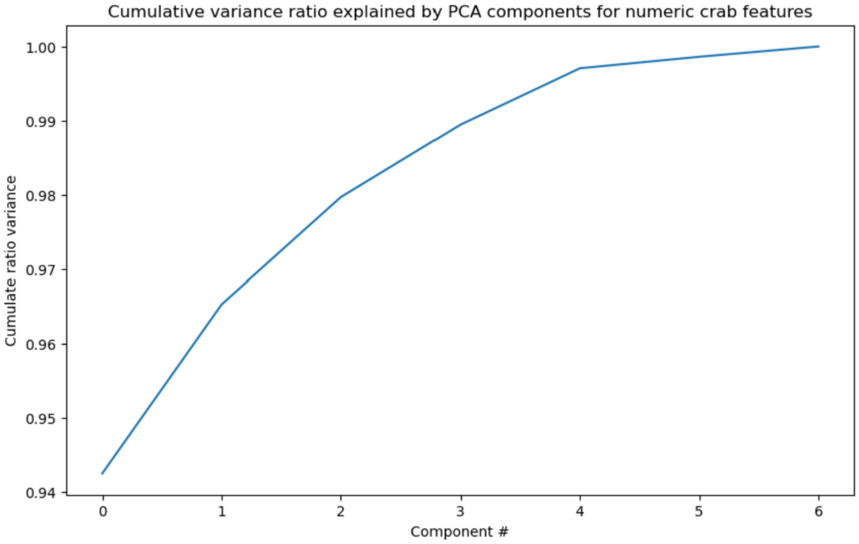

Figure 5. Cumulative explained variance ratio for vs number of principal components for principal component analysis completed on scaled crab age numeric feature set.

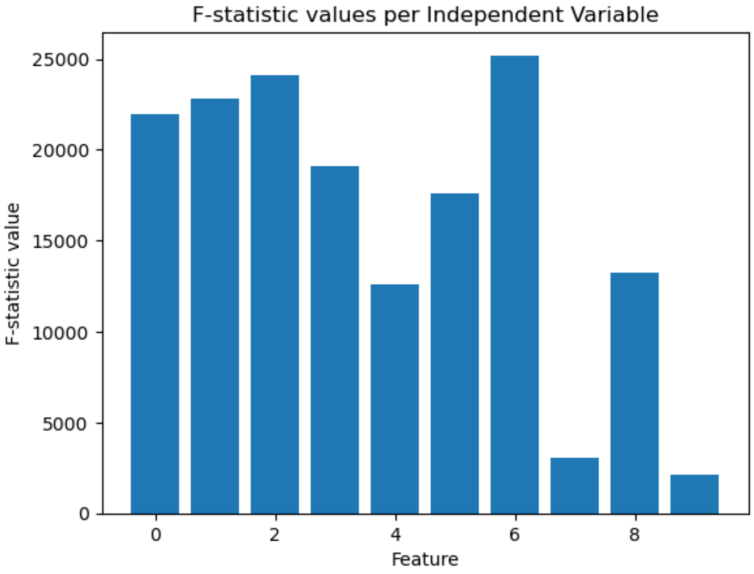

Figure 6. Feature selection performed prior to linear regression modeling.

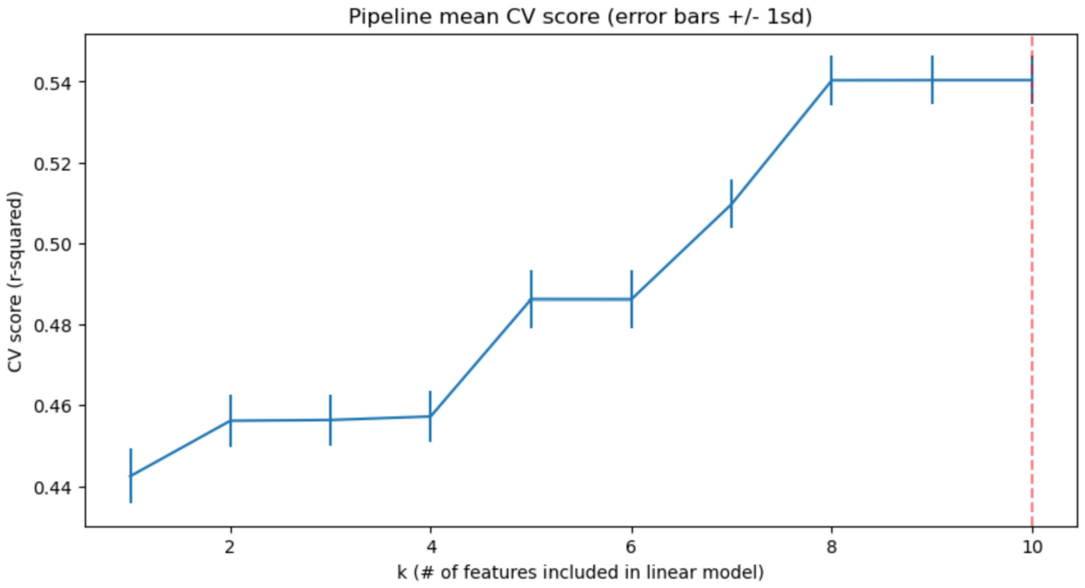

Figure 7. 5-fold cross validation, coefficient of determination scores for linear regression models created with each possible number of features in the dataset.

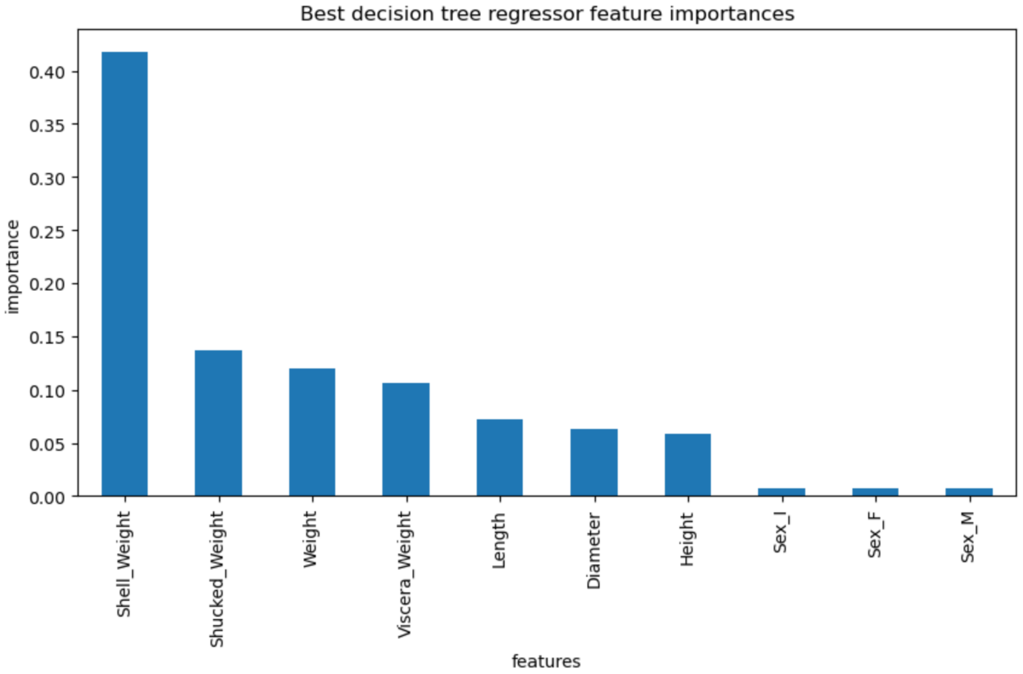

Figure 8. Decision-tree default parameter model's feature importances. Weight-related attributes were ranked to be those with highest importances.