Capstone 2: Crab Age Presentation

Jennifer Markham

Problem Identification

- 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
- After a certain age, crab growth becomes limited and any size gains become negligible
- To reduce cost and increase profits, it is critical for commercial crab farmers to know the optimal age to harvest crabs
- The goal is to predict crab based on physical attributes to help optimize crab farming practices

Recommendations/Key Findings

- A Decision-Tree Regressor model using default parameters most accurately predicted crab age

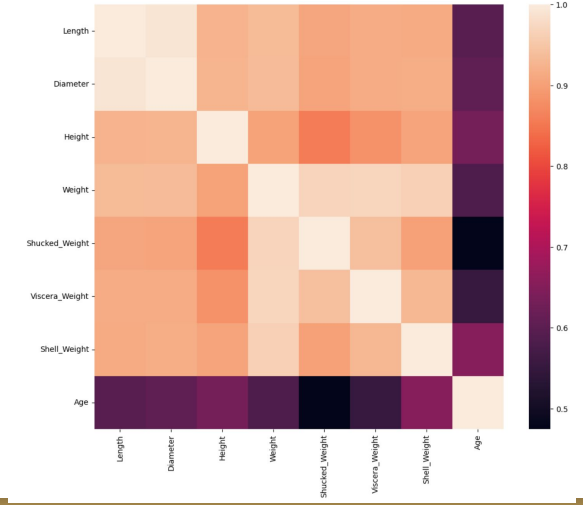
- The physical features of highest importance in this model were the weight-related attributes:
  - Shell weight
  - Shucked weight
  - Weight
  - Viscera weight

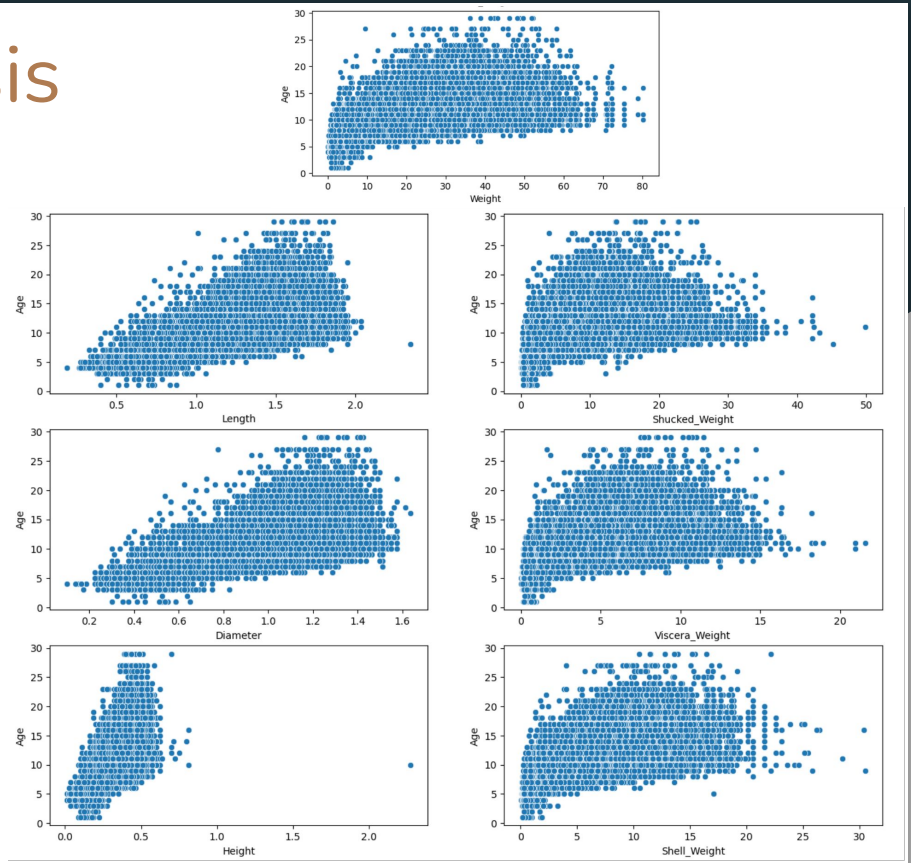
- Crab genders were the lowest importance features in this model

- 5-fold cross-validation using the default Decision-Tree Regressor model predicted crab age with a mean absolute error of ~0.119

Model Results/Analysis

- Scatterplots of Age vs each numeric attribute in the dataset were constructed to determine if any strong linear correlations existed for a single attribute
- A correlation heatmap was created to help determine if any numeric attributes were strongly correlated to crab age





Model Results/Analysis

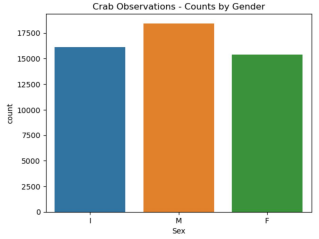
- Pearson correlation coefficients between Age & each numeric attribute were calculated to quantify any linear correlations, the strongest correlations were ~0.60-0.65 from the attributes below
    - Shell Weight
    - Height
    - Diameter

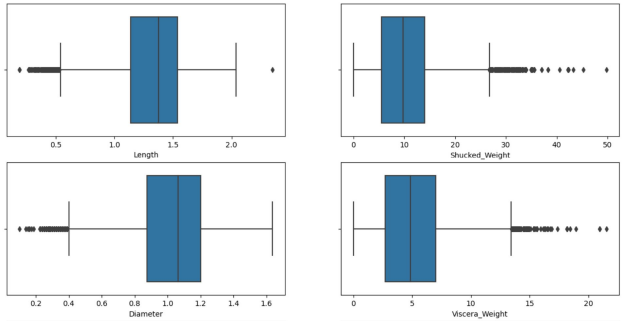
- Pearson correlation coefficients between Age & each numeric attribute were calculated on subsetted data by crab gender as well to quantify if any stronger linear correlations existed by Gender

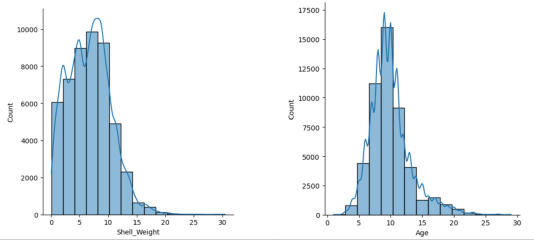
- Indeterminate gender data had the highest Pearson correlation coefficients between Age & numeric attributes, ranging from ~0.6-0.7 with the attributes below:
    - Shucked Weight
    - Viscera Weight
    - Shell Weight

Model Results/Analysis

- The distribution of each crab attribute was reviewed prior to modelling 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
- Each numeric attribute's distribution was reviewed using a boxplot & histogram
- Numeric attribute outliers were removed using the interquartile range approach described below
  - Upper bound for outliers: Q3 + 1.5 * IQR
  - Lower bound for outliers: Q1 - 1.5 * IQR



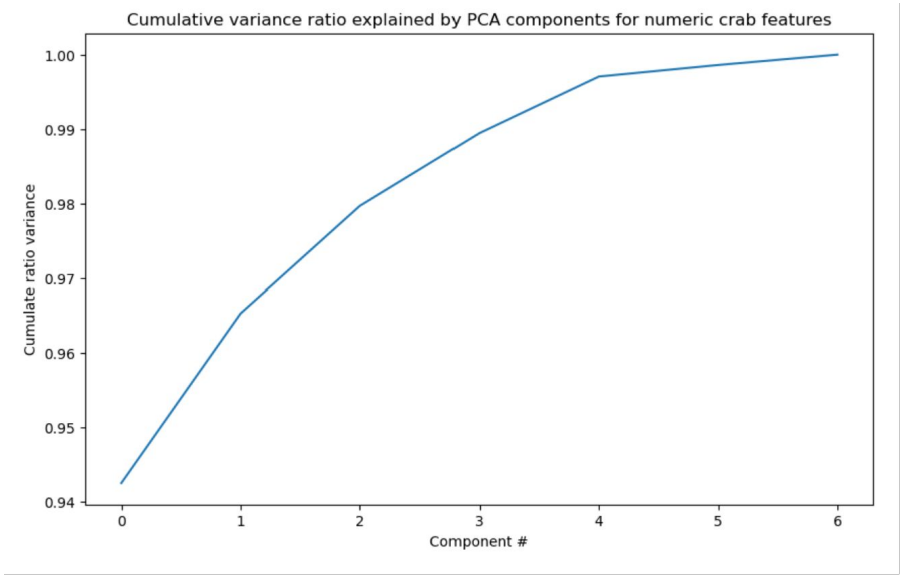




Model Results/Analysis

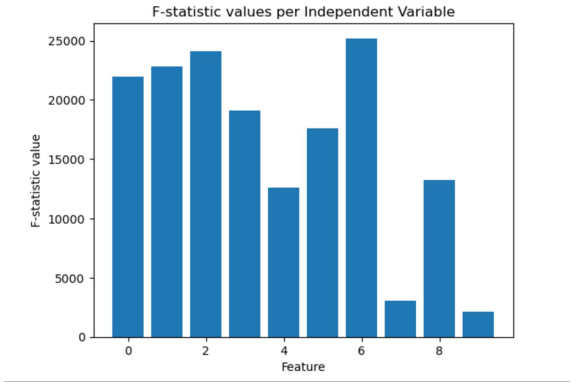
- A principal component analysis was completed on the scaled data with outliers removed to determine if dimensionality of the dataset could be successfully reduced

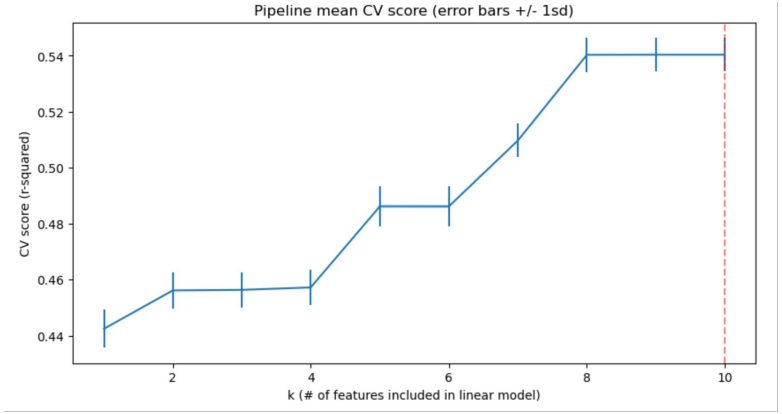
- The PCA analysis showed that dimensionality of the dataset could be reduced such that 2 principal components consisting of linear combinations of the crab features could account for ~97% of the cumulative variance seen in the dataset



Model Results/Analysis

- Feature selection was performed to determine which number of features should be included in the model
- Linear Regression modelling was subsequently performed, trialling inclusion of each number of possible features
- 5-fold cross validation was performed for each linear model created
- The best performance was achieved using a linear model including all features
  - Coefficient of determination: 0.54
  - Mean absolute error: 1.20
  - Mean square error: 2.47



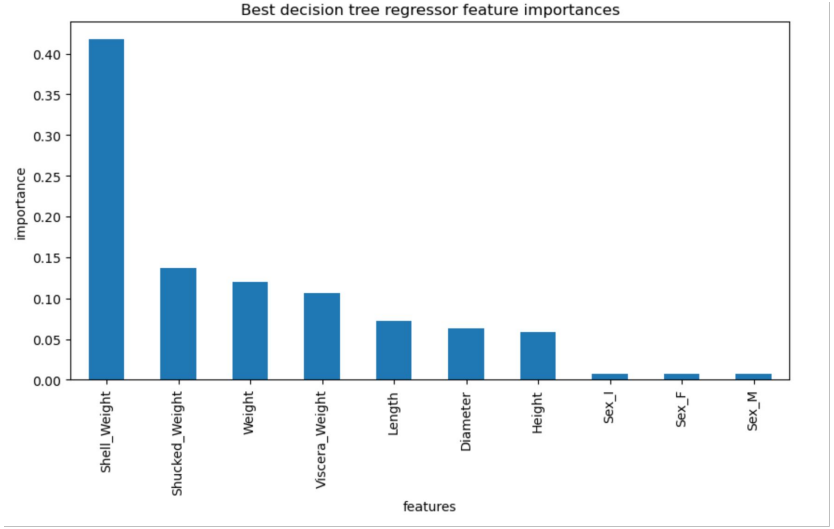


Model Results/Analysis

- Decision-Tree & Random-Forest Regressor models were also trialled using default parameters & hyperparameter tuning with RandomSearch

- Optimum performance was found using the default Decision-Tree Regressor model, with mean absolute error of 0.119

- The default Decision-Tree Regressor model listed the weight attributes below as those with the highest feature importances
  - Shell Weight
  - Shucked Weight
  - Weight
  - Viscera Weight



Summary/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