Predicting Crab Age with Machine Learning

Justin Heng

Graduate Student in Computer Science

California State University of Fullerton

Lakewood, California

justinheng@csu.fullerton.edu

Brian Lucero

Undergraduate Student in Computer Science

California State University of Fullerton

South El Monte, California

13rianlucero@csu.fullerton.edu

***Abstract*— Machine learning can be used to predict the age of crabs. It can be more accurate than simply weighing a crab to estimate its age. Several different models can be used, though support vector regression was found to be the most accurate in this experiment.**

I. Introduction

Crab is very tasty and many countries of the world import huge amounts of crabs for consumption every year. The main benefits of crab farming are, labor cost is very low, production cost is comparatively lower and they grow very fast. Commercial crab farming business is developing the lifestyle of the people of coastal areas. By proper care and management we can earn more from crab farming business than shrimp farming. You can raise mud crabs in two systems. Grow out farming and fattening systems. For a commercial crab farmer knowing the right age of the crab helps them decide if and when to harvest the crabs. Beyond a certain age, there is negligible growth in crab's physical characteristics and hence, it is important to time the harvesting to reduce cost and increase profit.

II. Background

According to the Chesapeake Bay Partners Program [3], blue crabs reach maturity in about 12 to 18 months. For simplicity's sake, it will be assumed that a crab will be mature and ready to harvest after 12 months in age. There are several other factors that contribute to the harvestability of crabs. Though length is usually the deciding factor for harvesting crabs, age will be used here since machine learning can be applied. There are also regulations that prevent certain crabs from being harvested. Namely, it is illegal to harvest female crabs that are holding eggs. This is another feature that will be ignored in this experiment.

III. Dataset and Data Preprocessing

The dataset that is being used was taken from Kaggle [1]. It contains over 1000 samples and nine features each. The features are "Sex", "Length", "Diameter", "Height", "Weight", "Shucked Weight", "Viscera Weight", "Shell Weight", and "Age". Fortunately, all of the data was present and no values were missing. In the case that values were missing, that specific data point could be taken out in order to avoid any errors during calculations. Since "Sex" had a value of either "M" for male, "F" for female, and "I" for indeterminate, conversions were necessary in order to give the feature a numerical value. Male was given a numerical value of 1, female was given 2, and indeterminate was given 1.5. These values were stored into a new feature called "SexValue".

To perform feature selection, the Pearson correlation coefficient was found for each of the eight values in relation to "Age". The results are in the table below.

SexValue	0.0337
Length	0.555
Diameter	0.574
Height	0.552
Weight	0.539
Shucked Weight	0.419
Viscera Weight	0.501
Shell Weight	0.625

Table 1. Pearson correlation coefficients

Sex appears to have little to no correlation to age which is as expected. All the other features have a very high Pearson correlation coefficient meaning age is highly correlated to these features. Shell weight has the highest correlation with age at 0.625. All of the features will be used in the prediction models with the exception of "SexValue" since there is no correlation. The models will also be tested using only "Weight" as a feature as it is more practical than measuring a crab for seven other features.

The dataset is split into a training set and a test set using the test_train_split function from the Scikit Learn library. The sets are labeled as "X_train", "y_train", "X_test", and "y_test" respectively. The training set consists of 70% of the data and the test set contains 30%. A random seed of 132 is used in order to randomly split the dataset properly.

IV. Methodology

The crab age prediction is done using three different models in order to get a wide range of results. In doing so, the model that performs the best can be found and used in future crab age predictions. The three models are K-nearest neighbor, linear regression, and support vector regression. The models are built in Python using third party libraries. The algorithms for K-nearest neighbor, multiple linear regression, and support vector regression are part of the Scikit Learn library [2] for Python.

A. K-Nearest Neighbor

K-nearest neighbor is the first model that is implemented. In order to find a good value for k, the model is run with a k value

between 1 and 30. The following graph shows the average error with each corresponding value of k.

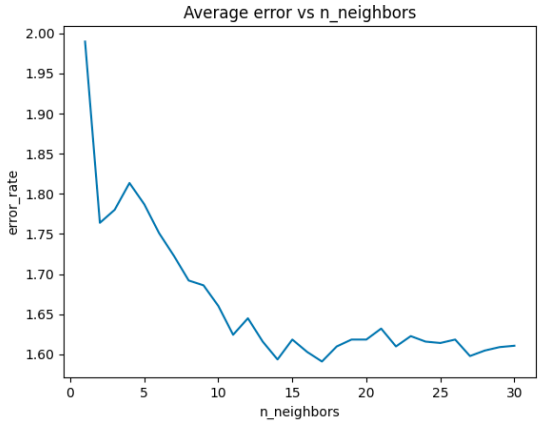

Fig. 1. Average error versus n_neighbors

The average error is greatest when k is equal to 1 and starts to taper off around k equal to 15. For the rest of this experiment, k will be set to 20 since the error rate appears to stabilize around this value.

In order to visually gauge how the model performs, the first 50 values that are predicted will be plotted alongside the actual values. The average error in months will also be taken.

B. Linear Regression

The Multiple Linear Regression model is intended to have various data preprocessing activities included in the overall dataset configuration workflow. This includes building the data frame variable and using both train-test split and cross validation as two main branches of Multiple Linear Regression predictions. Using Scikit Learn, we can build polynomial features which would construct a curve instead of a line using polynomial regression. This is to fit polynomials of varying degrees onto the initial dataset in order to fit the curve according to the feature values plot, in order to avoid making the model “see” the test data

C. Support Vector Regression

Support vector regression is a modified form of support vector machine that is designed to work with regression problems instead of classification problems. The model is straightforward as the only arguments it takes are the training sets and the x-values for the test set.. It is fitted using the training set and creates predictions using the test set. The first 50 values will also be plotted alongside the actual values and the average error will be found.

V. Results

In order to have a baseline to compare to, simple linear regression is used to find the age of a crab using only its weight. The graph shows the output. The average error is 1.939 months.

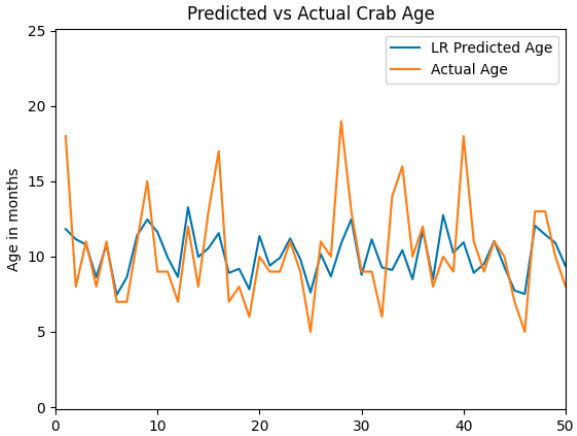

Fig. 2. Predicting age using only weight

The following graph shows all the models' predictions alongside the actual age of the crab for the first 50 data points.

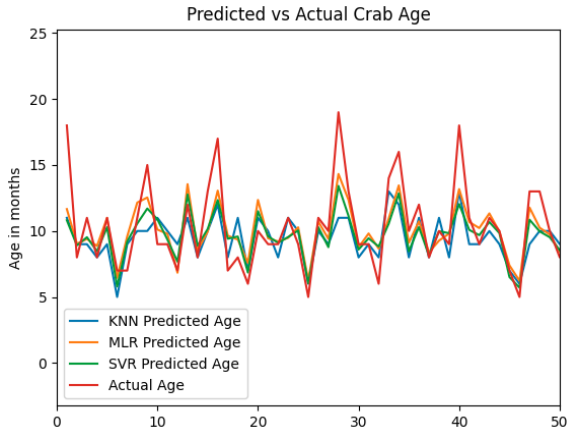

Fig. 3. Results of all models

All the models performed well with only small differences between each. The models all seemed to struggle with predicting older crabs and tended to underestimate the age.

A. K-Nearest Neighbor

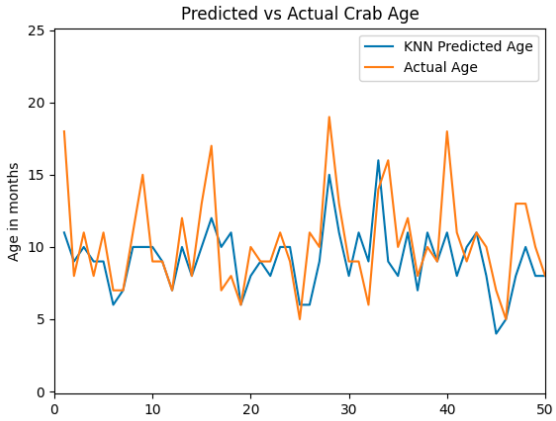

Fig. 4. Results of KNN prediction model

The average error between the predicted age and the actual age was about 1.610 months. Although the amount of error is small, the K-nearest neighbor model actually performed the worst out of the three by a small margin. Unlike the other two models, it was not as accurate at predicting ages under 12 months. For fully matured crabs, it performed average at predicting its age. Overall, it is a decent and usable model.

B. Linear Regression

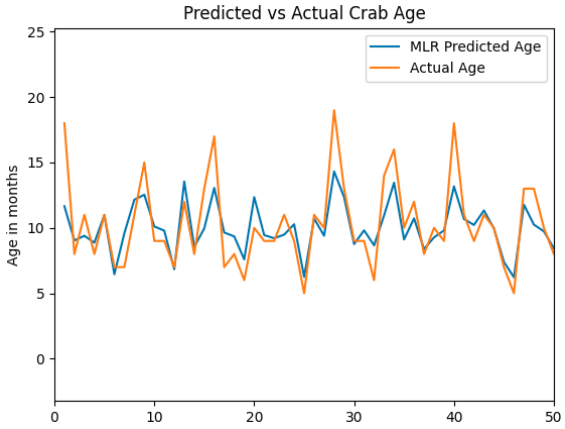

Fig. 5. Results of Multiple Linear Regression prediction model

The average error between the predicted age and the actual age was about 1.560 months. The Multiple Linear Regression model fared well in its performance compared to the previous KNN model implementation. For the sake of time, we could only perform analysis using all of the same features that the other models use, simply excluding the gender of the crabs. So, the MLR model is built plain without much feature filtering nor feature selection using cross validation. However, the model also managed to slightly out-perform the former KNN model by a small measure. Although MLR models with a vanilla design tend to suffer in performance from overfitting, we observe the current set of features to work similarly well between the two initial models, KNN and MLR. The model's performance metrics were tested by calculating the average error

and that served as the main classifier evaluation benchmark.

C. Support Vector Regression

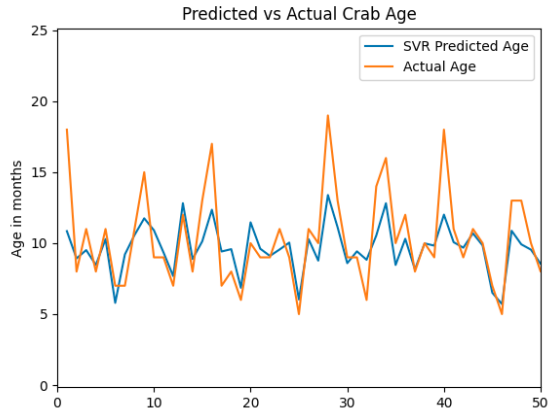

Fig. 6. Results of SVR prediction model

The average error between the predicted age and the actual age was about 1.471 months. The support vector regression model overall had the least amount of error. Multiple linear regression was better at predicting mature crab ages past 12 months, but support vector regression was much more accurate at predicting crabs that were under 12 months.

VI. Conclusion

Overall, the models were able to predict the age of crabs reasonably well. On average, the predictions were off by about 1.5 months. Although support vector regression performed slightly better than the other two models, it was still close enough that any of the models could be used with satisfactory results. Multiple linear regression was found to be slightly better at predicting older crabs while support vector regression was better at predicting younger crabs. K-nearest neighbor was average overall. What is important to note is that the predictions for all three models were more accurate when the age of the crab was less than 12 months. This makes sense because after a crab reaches full maturity around 12 months, its growth comes to a halt and it is harder to predict its age since its features stay roughly the same. Therefore, predicting the age of a crab becomes less accurate the longer a crab has matured.

References

[1] https://www.kaggle.com/datasets/sidhus/crab-age-prediction

[2] https://scikit-learn.org/stable/modules/svm.html

[3] https://repository.library.noaa.gov/view/noaa/16273/noaa_16273_DS4.pdf

[4] https://faculty.math.illinois.edu/~hildebr/tex/latex-start.html

[5] https://github.com/krishnaik06/Multiple-Linear-Regression

[6] https://github.com/13rianlucero/CrabAgePrediction