Predicting Crab Age with Machine Learning

CPSC-483 Machine Learning

🦀 Justin Heng, Brian Lucero 🦀

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.

Introduction

The Problem ✓	Why it's important? ✓	Our Solution Strategy ✓
It is quite difficult to determine a crab's age due to their molting cycles which happen throughout their whole life. Essentially, the failure to harvest at an ideal age, increases cost and crab lives go to waste.	*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.*	Prepare crab data and use it to train several machine learning models. Thus, given certain physcial chraracteristics and the corresponding values, the ML models will accurately determine the age of the crabs.

Background

- Assume that a crab is mature and ready to harvest after 12 months
- Ignore other features that affect a crab's harvestability such as egg laying crabs
- Predict age rather than weight since machine learning is more applicable (pointless to predict weight)

Dataset

- dataset from Kaggle
- over 1000 samples and nine features each
    - "Sex", "Length", "Diameter", "Height", "Weight", "Shucked Weight", "Viscera Weight", "Shell Weight", and "Age"

	A	B	C	D	E	F	G	H	I
1	Sex	Length	Diameter	Height	Weight	Shucked V	Viscera W	Shell Wei	Age
2	F	1.4375	1.175	0.4125	24.63572	12.33203	5.584852	6.747181	9
3	M	0.8875	0.65	0.2125	5.40058	2.29631	1.374951	1.559223	6
4	I	1.0375	0.775	0.25	7.952035	3.231843	1.601747	2.764076	6
5	F	1.175	0.8875	0.25	13.48019	4.748541	2.282135	5.244658	10
6	I	0.8875	0.6625	0.2125	6.903103	3.458639	1.488349	1.70097	6
7	F	1.55	1.1625	0.35	28.66134	13.57941	6.761356	7.229123	8
8	F	1.3	1	0.325	17.70426	6.095143	5.854172	4.819415	15
9	M	1.325	1.0125	0.375	23.57261	9.979024	5.301357	7.158249	10
10	I	1.5875	1.25	0.4125	42.21241	20.26989	9.766403	10.24834	13

Data Preprocessing

- Converting sex to numerical values
  - Male = 1
  - Female = 2
  - Indeterminate = 1.5
- Train test split
  - ```
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=132)
    ```
    ```

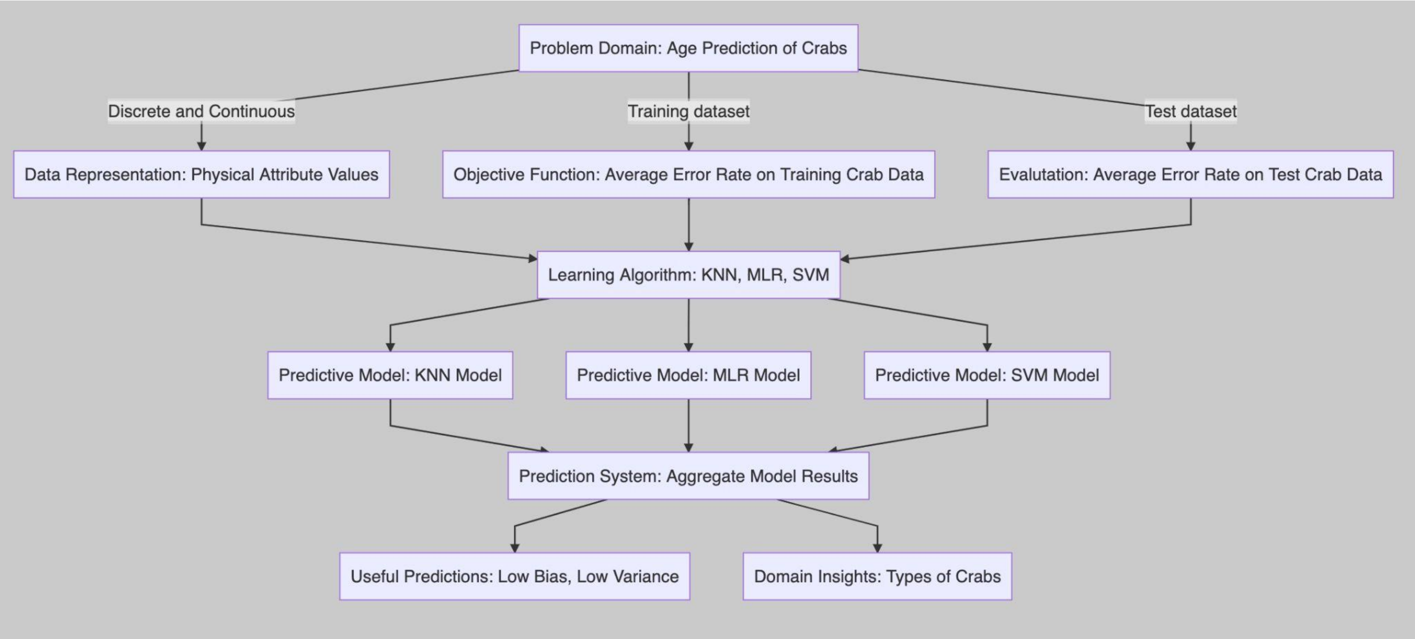
Feature Selection

- Pearson correlation coefficients
- Ignore sex and keep all the other features
- > 0.5 is very high correlation

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

Methodology



Methodology

- 3 machine learning models and 1 baseline model
  - Simple linear regression of Weight vs Age (baseline)
  - K-nearest neighbor (ML)
  - Multiple Linear Regression (ML)
  - Support Vector Regression (ML)
- Scikit Learn Python libraries

```
#KNN
neigh = KNeighborsClassifier(n_neighbors=20)
neigh.fit(X_train, numpy.ravel(y_train))
knn_predict = neigh.predict(X_test)

#Multiple Linear Regression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

y_pred = regressor.predict(X_test)

score = r2_score(y_test,y_pred)

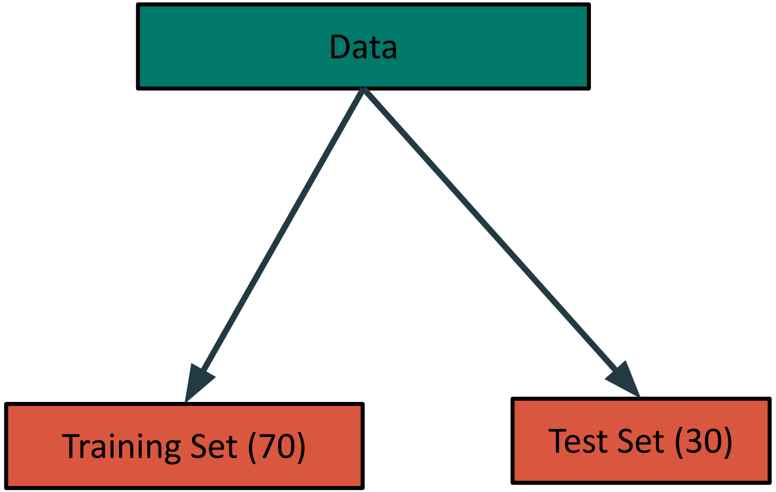
#Linear regression with weight
regressor2 = LinearRegression()
regressor2.fit(numpy.array(X_train["Weight"]).reshape((-1,1)), y_train)

y_pred2 = regressor2.predict(numpy.array(X_test["Weight"]).reshape((-1,1)))


#SVR
regr = svm.SVR()
regr.fit(X_train, numpy.ravel(y_train))
regr_predict = regr.predict(X_test)
```

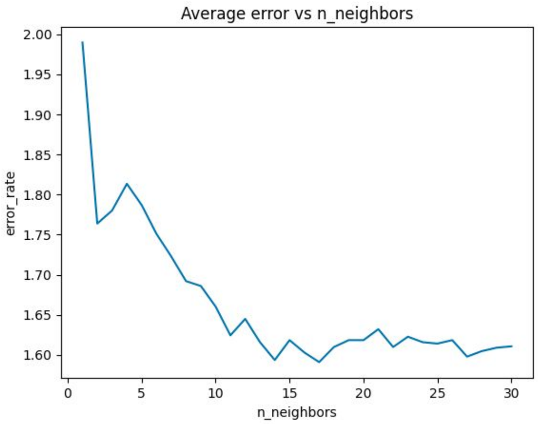
Methodology (MLR)

- Feature Selection
  - Removed gender
- Train-Test Split
- Train model
- Evaluate results
- **Measure Average Error**

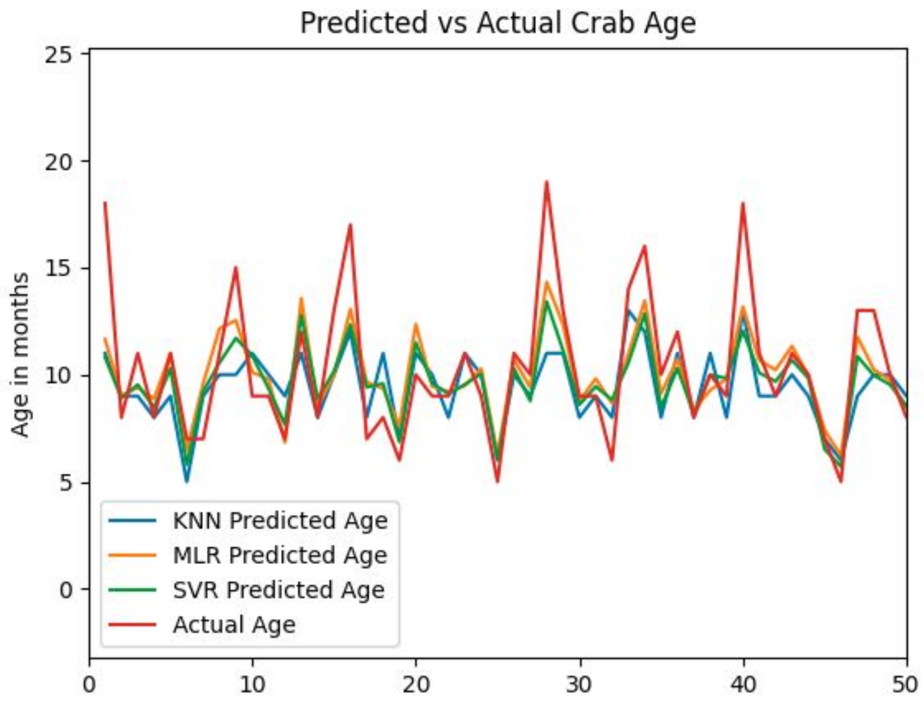


Methodology (Knn)

- Need to find what value to use for k
- Average error for 0 < k < 30
- We will use k = 20

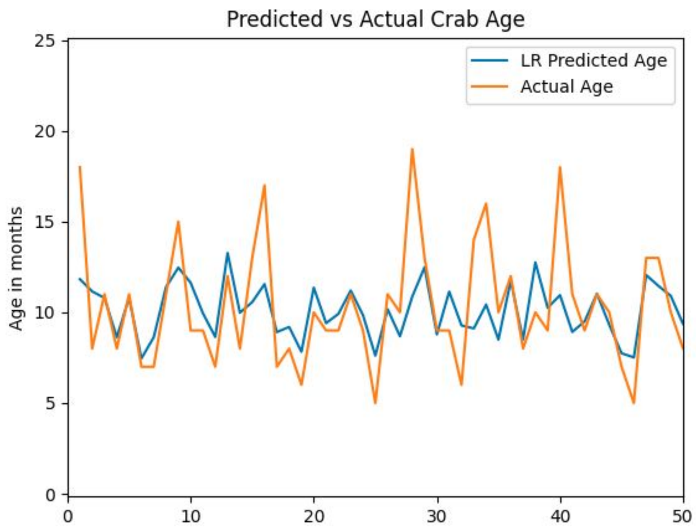


Results


Predicted vs Actual Crab Age
Age in months
25
20
15
10
5
0
0
10
20
30
40
50
KNN Predicted Age
MLR Predicted Age
SVR Predicted Age
Actual Age

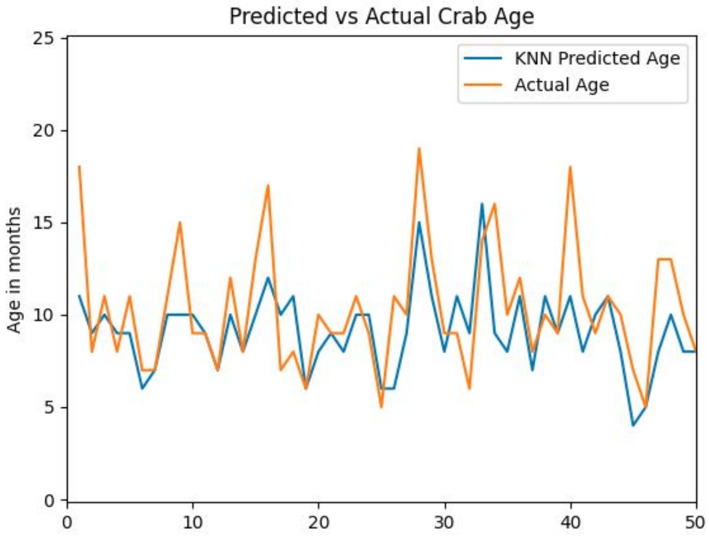
Results (Baseline)

- Linear regression
- Only uses weight to determine age
- Average error: 1.9 months



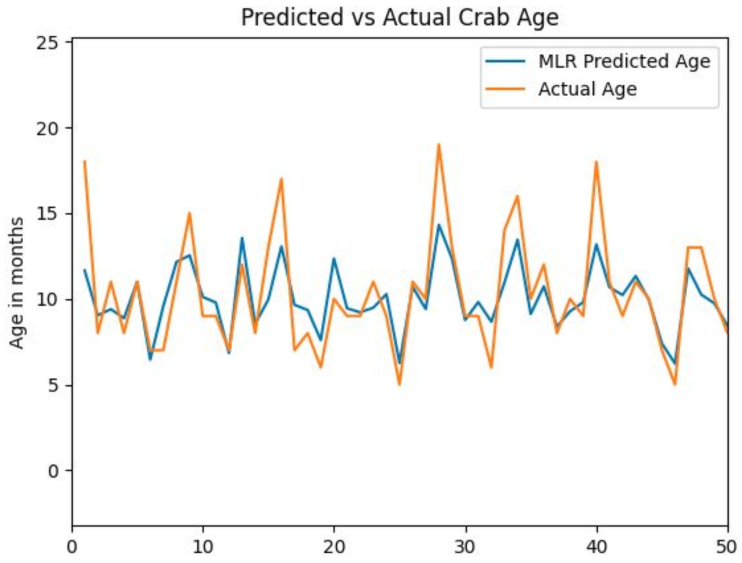
Results (Knn)

- Outperformed the baseline
- Slightly the worst out of the three ML models
- Not as accurate at predicting ages under 12 months
- Average error: 1.6 months



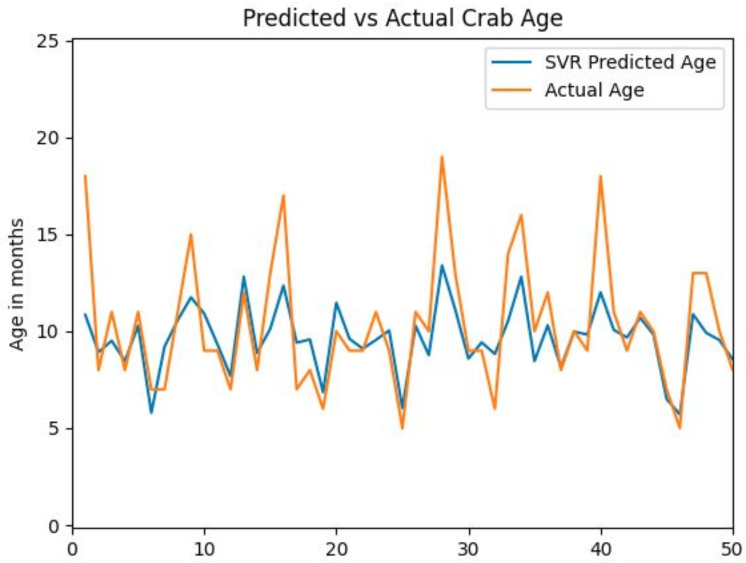
Results (Multiple Linear Regression)

- Outperformed the baseline
- Better at predicting mature ages above 12 months
- Overall, a good model
- Average error: 1.5 months



Results (Support Vector Regression)

- Outperformed the baseline
- Accurate at predicting ages under 12 months
- Had the least amount of error
- Average error: 1.4 months



Results

Model	Type	Error (months)
Linear Regression (Weight vs Age)	Baseline	1.939
K-nearest Neighbor	ML	1.610
Multiple Linear Regression	ML	1.560
Support Vector Regression	ML	1.471

Conclusion

- Machine learning outperformed simple linear regression
- On average, the models had an error of about 1.5 months compared to 2.0 months
- Support vector regression had a slight lead, but multiple linear regression and k-nearest neighbor were good predictors as well
- Predictions were good up until 12 months when the crabs reached full maturity

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