Predicting the Age of Abolone Using Physical Characteristics

Quinn Lennemann
University of Nebraska - Lincoln

Kolby Johnson
University of Nebraska - Lincoln

Linhan Li
University of Nebraska - Lincoln

***Abstract*—The age of abalone is normally found by counting the rings on the shell, but this can also be done by analyzing physical characteristics of the animal. We propose using three different regression models on our dataset, and also seeing how principal component analysis (PCA) affects the speed and accuracy of our resuts. We found that our best model is able to predict ages with a mean error of 1.32 years, and that PCA is able to both reduce the speed and increase the accuracy of our models.**

I. Introduction

Abalone are a type of sea snail which are commonly used for food. They can be found in cold coasts around the world. In North America, they can be found along much of the Pacific Coast. They are able to live up to 50 years. Different species of abalone can be found at different depths [1]. Although they were once extremely plentiful along the coast, their populations have taken a sharp decline in recent years due to overharvesting and disease. They are now a protected species, and harvesting abalone is illegal in some protected locations.

The age of an abaolone can be determined by the number of rings on its shell in a method similar to reading the number of rings that a tree stump has. While this may be useful in gathering data to assist in revitalizing abalone populations, it is also very time consuming. However, machine learning may be able to help, as it could be possible to get an accurate estimate of an abalone's age using measurements that are much easier to obtain.

II. Problem Definition

In this project, we will be trying to determine the age of abalone using their physical characteristics. These measurements include sex, length, diameter height, and four various weight attributes. This task will be performed using regression, as the age of an abalone can be represented using a numerical value.

As mentioned before, this problem is important because the abalone population is in decline, especially off the west coast of California. One of the group members lives in an area that has an abalone population. Many people dive in the area to hunt for abalone, and many nearby areas are named after the animal. In fact, his grandfather often searched for abalone along the shore. Therefore, we thought it would be interesting to do a project related to the animal.

Another reason for picking this project was more practical. Since this group is comprised of undergraduate students, we thought that it would be best to use a dataset that we would know a good solution for, and this dataset seems like it would be a good fit for a regression model.

III. Dataset

This dataset was created during a 1994 study of abalone off the coast of Tasmania [2]. It was not originally intended to be used to supplement machine learning, however, this dataset is excellent for the task. This dataset has been modified to remove missing values. Additionally, all of the continous features have been scaled down by a factor of 200. For the purposes of this project, we did not find it necessary to scale them up back to their original values.

This dataset can be found at the following link: "https://archive.ics.uci.edu/ml/datasets/Abalone".

The dataset contains 4,177 rows of data, and each instance has 8 dimensions. Those dimensions include sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and number of rings. All of the dimensions excluding sex and number of rings are continuous variables. The distance features were measured in millimeters, and the weight features were measured in grams. The sex column contains "M' for male, "F" for female, and "I" for infant. The number of rings is represented by discrete integer values.

We used one-hot encoding for the sex feature, creating columns for "M", "F", and "I", with a 1 in the feature which correctly describes the instance and a 0 in the other feature columns.

Additionally, we recoded the target vector from representing the number of rings to representing the age of the abalone in years. 1.5 was added to every feature in the rings feature column, creating a new target vector. This was treated as a continuous feature.

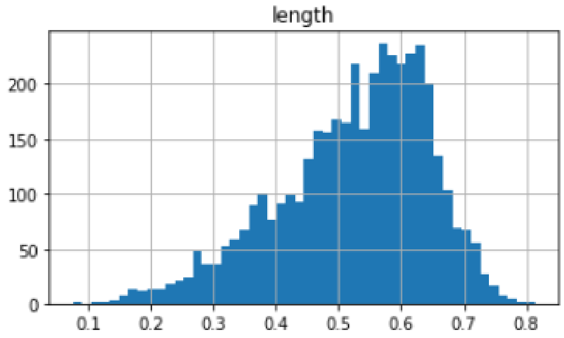
IV. Data Preprocessing

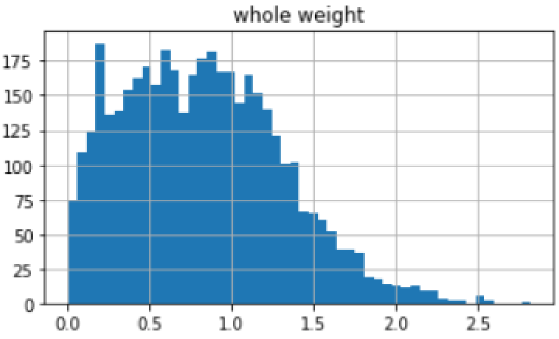
We started our preprocessing by creating columns for ratios of the weight of specific parts of the abalone compared to the total weight. We then graph the distribution of the data with the new columns to look for outliers. We find the outliers and drop them to make our data more precise. After that, we drop the unwanted/unneeded columns like sex and the ratio of weights columns to clean up the data a little bit.

Finally we use principle component analysis (PCA) on the data to reduce the dimensionality. Using PCA, we were able to drop the number of remaining features from 8 to 4 while still keeping the vast majority of the variance, at 99.718%. This is useful, as models that use PCA will be able to run faster, and they may even have improved accuracy.

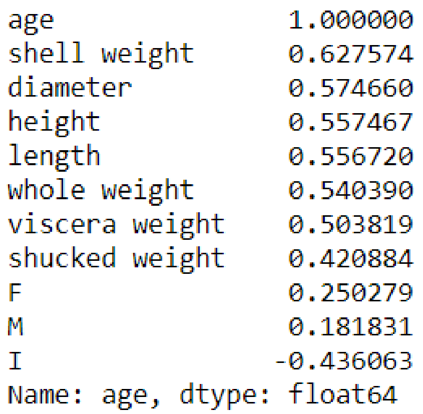
V. Exploratory Data Analysis

We performed multiple methods of data analysis on our dataset. The first thing we did was look at the distribution of our data. We noticed that many features of our dataset had a skewed normal distribution. The length features were skewed towards higher values, and the weight features were skewed towards lower values. This led to us using aggressive principal component analysis on the data, as we could see that several features shared similar variance with each other.





Another analysis we performed on the data involves checking the correlation of each feature when compared to age. We noticed an interesting point with the sex features. Both the male and female features had a low positive correlation with the age, with ”F” having a correlation of 0.25 and ”M” having a correlation of 0.18. However, the infants has a significant negative correlation with age, at -0.44. This showed us that while the gender of the abalone is not important, the infancy of an abalone is very important.



```
age               1.000000
shell weight      0.627574
diameter          0.574660
height            0.557467
length            0.556720
whole weight      0.540390
viscera weight    0.503819
shucked weight    0.420884
F                 0.250279
M                 0.181831
I                -0.436063
Name: age, dtype: float64
```

The following tables provide the mean, standard deviation, and quartiles for each variable in the dataset. The target variable has been bolded.

	M	F	I
mean	0.366	0.313	0.321
std	0.482	0.464	0.467
25%	0	0	0
50%	0	0	0
75%	1	1	1

	length	diameter	height
mean	0.524	0.408	0.140
std	0.12	0.010	0.042
25%	0.45	0.055	0.115
50%	0.55	0.35	0.140
75%	0.62	0.425	0.165

	whole weight	shucked weight	viscera weight
mean	0.829	0.359	0.181
std	0.222	0.222	0.110
25%	0.442	0.186	0.094
50%	0.800	0.336	0.171
75%	1.153	0.502	0.253

	shell weight	**age**
mean	0.239	**11.434**
std	0.139	**3.224**
25%	0.13	**9.5**
50%	0.234	**10.5**
75%	0.329	**12.5**

VI. Evaluation Metric

We used two different evaluation metrics for our project. First, we looked at mean absolute error (MAE). Although we have been using mean squared error (MSE) for much of this course, we decided to use MAE instead for two main reasons. The first reason is that MAE is much easier to understand at a glance. For our project, it is simply the average difference between the true age and the age that our models predict. Intuitively understanding MSE is much harder, as it doesn’t have an exact unit. Secondly, we felt safe using MAE because we don’t have a large number of outliers in our dataset. MSE is good for emphasizing the error of large outliers, but since this is a dataset that has already been cleaned, this is not as important.

The second evaluation metric we used is r2 score. This is a very common metric to use for regression models, as it defines how well the model fits to the dataset. Usually, a higher r2 score means that the model will be more accurate, as it shows that the model is making predictions very close to the correct values.

VII. METHODS

A. Bayesian Ridge Regression

Bayesian ridge regression is a model defined in probabilistic terms, that uses explicit priors in the parameters. It also uses L2 norm for regularization and with using laplace priors as coefficients it is the same as L1 regularization. Essentially, it is the ordinary least squares (OLS) solution, but it also uses penalty to avoid overfitting. We decided to use this model since it is good on datasets that have collinearity among the dimensions, which our dataset has, since the weight measurements are all very similar to each other.

Since bayesian ridge regression is fast on our dataset, we were able to tune many hyperparameters. The parameter ”n_iter” used [1, 5, 100], ”alpha_1” used [0.000001, 0.0001], ”alpha_2” used [1, 5, 500], ”lambda_1” used [0.1, 0.5, 100], and ”lambda_2” used [0.00000001, 0.000001, 0.0001]. Our optimal hyperparameters were 0.00001 for ”alpha_1”, 500 for ”alpha_2”, 0.5 for ”lambda_1”, 0.00000001 for ”lambda_2, and 1 for ”n_iter”. The hyperparameters remained the same between non-PCA and PCA variants.

B. Stochastic Gradient Descent

Stochastic gradient descent picks a random row in the training set for every step, and the gradients are only computed for that single row. This makes the algorithm fast since only a small amount of data is required for each iteration. Since it only looks at a small amount of data, there only needs to be one row in memory allowing training on very large data sets. With the algorithm randomly selecting rows at each iteration it makes getting out of local minimum easier and giving you a better chance of reaching the global minimum. Our data set isn’t incredibly large but we chose this algorithm because we wanted to see how it may work on a large data set and for the speed boost it can offer.

We tuned four hyperparameters. The parameter ”alpha” used [0.001, .00001, .000001], ”l1_ratio” used [0.5, 0.2, 0.1], ”max_iter” used [15000, 20000, 30000], and ”eta0” used [0.1, 0.01, 0.001]. Additionally, we set our penalty to ”elasticnet”, which allowed for the ”l1_ratio” to be tuned. The optimal hyperparameters were 0.000001 for ”alpha”, 0.1 for ”eta0”, 0.1 for ”l1_ratio” and 30000 for ”max_iter”. The only difference between the non-PCA and PCA models was that the optimal ”max_iter” for the PCA model was lowered to 15000.

C. Random Forest Regression

Random forest regression uses a group of deep decision trees with each tree using a different random subset of the training data. Using deep decision trees allows us to fit complex data sets. Since random forest has each tree on a different random subset it reduces the variance, therefore boosting the performance of the model. With each of the trees being randomly different from each other it causes decorrelation between each tree predictions therefore improving generalization. We chose this algorithm because we have a somewhat complex data set with a high variance, and random forest regression helps with both of those things.

Since random forest regression can be quite slow, we only tuned a few hyperparameters. The values we used for ”n_estimators” were [50, 500, 1500], and the values we used for ”min_samples_split” were [10, 25, 50]. Our optimal hyperparameters were 500 and 50, respectively. The hyperparameters remained the same between non-PCA and PCA variants.

VIII. SUMMARY OF RESULTS

A. Bayesian Ridge Regression

Bayesian ridge regression was by far the fastest out of all of the models we used, as the average time for a fit with the standard model was 4.2 ms, while the average time for a fit with the PCA model was only 0.5 ms. However, this speed has a disadvantage, as it was also the model that performed the worst, although it wasn’t extremely significant.

However, we noticed that just because an algorithm is faster, it doesn’t always mean that it is more accurate. This is evidenced by the fact that the PCA variant was more accurate despite being much faster. This is due to the fact that PCA can sometimes better characterize the variance of the data than the original ”full” dataset can.

Bayesian Ridge Regression		
PCA	Without	With
MAE	1.458	1.415
r2 Score	0.485	0.526

B. Stochastic Gradient Descent

Our stochastic gradient descent model was also quite speedy, although it was still slightly slower than the bayesian ridge regression model. The non-PCA model had an average fit time of 8.9 ms, and the PCA model had an average fit of 5.2 ms. The only significant performance boost this model had over the ridge regression model was that the MAE with PCA was 0.033 higher, which still isn’t much. It had worse performance in all of the other categories.

This is likely due to the fact that the dataset we are using may be relatively big compared to other datasets we have used in the course, but it may not be big enough to outweigh the downsides of stochastic gradient descent.

Stochastic Gradient Descent		
PCA	Without	With
MAE	1.491	1.409
r2 Score	0.477	0.523

C. Random Forest Regression

Finally, we will talk about our Random Forest Regression model. This model far outperformed our other models, with a MAE as low as 1.32 and with the only r2 score that nearly reaches 0.6. However, this model was also much slower than our other models, as it had a non-PCA fit time of 15.6 ms and a PCA fit time of 13.3 ms.

As mentioned previously, we think that this model performed the best on our data due to the high complexity of our dataset.

Random Forest Regression		
PCA	Without	With
MAE	1.383	1.318
r2 Score	0.520	0.586

Overall, we were quite pleased with our results, and didn't end up running into any major problems along the way.

IX. Conclusion & Future Work

In conclusion, we were able determine the age of abalone using their physical characteristics fairly accurately, with our highest mean absolute error being only 1.491. We found that the best model for our data set was random forest regression with a mean absolute error of 1.318 using PCA on the data. The use of PCA on our data reduced the dimensionality of our data by half while also reducing the error of our models. This was a very satisfactory result. In the future, we could use more models to see if there is improvement. Another change we may be able to make in the future would be to run our models on a more powerful machine, as the power of a standard home PC limited our ability to do extensive hyperparameter testing on our models.

References

[1] MARINe. 2020. Haliotis Rufescens. [online] Available at: https://marine.ucsc.edu/target/target-species-haliotis-rufescens.html.

[2] Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) ”The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait”, Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288